Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics

Subazini Thankaswamy-Kosalai, Partho Sen, Intawat Nookaew

PII: S0888-7543(17)30020-4
DOI: doi: 10.1016/j.ygeno.2017.03.001
Reference: YGENO 8867

To appear in: *Genomics*

Received date: 4 August 2016
Revised date: 7 March 2017
Accepted date: 8 March 2017

Please cite this article as: Subazini Thankaswamy-Kosalai, Partho Sen, Intawat Nookaew , Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Ygeno(2017), doi: 10.1016/j.ygeno.2017.03.001

Evaluation and assessment of read-mapping by multiple Next-generation sequencing aligners based on genome-wide characteristics

Subazini Thankaswamy-Kosalai[1#], Partho Sen[1#], Intawat Nookaew[1, 2*]

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96, Göteborg, Sweden.

[2] Comparative Genomics group, Bioscience Division, Oak ridge National Laboratory, Oak ridge, Tennessee 37831, USA.

[#] The authors contributed equally

[*]Corresponding Author

Email: intawat@chalmers.se , nookaewi@ornl.gov

Abstract

Massive data produced due to the advent of next-generation sequencing (NGS) technology is widely used for biological researches and medical diagnosis. The crucial step in NGS analysis is read alignment or mapping which is computationally intensive and complex. The mapping bias tends to affect the downstream analysis, including detection of polymorphisms. In order to provide guidelines to the biologist for suitable selection of aligners; we have evaluated and benchmarked 5 different aligners (BWA, Bowtie2, NovoAlign, Smalt and Stampy) and their mapping bias based on characteristics of 5 microbial genomes. Two million simulated read pairs of various size (36bp, 50bp, 72bp, 100bp, 125bp, 150bp, 200bp, 250bp and 300bp) were aligned. Specific alignment features such as sensitivity of mapping, percentage of properly paired reads, alignment time and effect of tandem repeats on incorrectly mapped reads were evaluated. BWA showed faster alignment followed by Bowtie2 and Smalt. NovoAlign and Stampy were comparatively slower. Most of the aligners showed high sensitivity towards long reads (> 100bp) mapping. On the other hand NovoAlign showed higher sensitivity towards both short reads (36bp, 50bp, 72bp) and long reads (> 100bp) mappings; It also showed higher sensitivity towards mapping a complex genome such as *Plasmodium falciparum*. The percentage of properly paired reads aligned by NovoAlign, BWA and Stampy were markedly higher. None of the aligners outperforms the others in the benchmark. We expect that the results from this study will be useful for the end user to choose aligner, thus enhance the accuracy of read mapping.

Keywords: Next-generation sequencing, NGS, aligners, mappers, alignments, mapping, algorithm, reads, genome.

1. Introduction

Recent advances in Next-Generation sequencing (NGS) technology generates huge amount of data. Proper handling and analysis of these data is necessary to answer biological questions [1]. NGS analysis has been widely applied in various genomics and transcriptomics studies [2]. The fundamental step in NGS analysis is mapping or alignment. Read mapping is the basis for further analysis; for instance, to estimate the abundance of transcripts and variant detection [3]. However, this step is biased by sequencing error and genome characteristics such as repetitive regions [4]. The reads may align in multiple location of the genome due to uncertainty in sequencing data [5]. Several aligners have been designed to attain high quality alignment with minimal mapping bias. The common mapping tools are based on hash table or index based algorithms [6]. The index-based aligners are slow, use more memory but appropriately maps long gaps. On the other hand heuristic based aligners are fast, uses less memory and can be used to map short reads [6]. Spliced aligners such as Tophat [7] and GEM [8] are most frequently used for aligning transcripts. BLAT [9], BWA [10], Bowtie2 [4] are used for aligning DNA sequences. BWA and Bowtie2 are index based aligners with Burrows Wheeler indexing algorithm.

Smalt (http://www.sanger.ac.uk/resources/software/SMALT/) and NovoAlign (http://www.novocraft.com) uses dynamic programming; Smalt is based on extension of the Smith Waterman algorithm. NovoAlign uses the Needleman Wunsch algorithm with affine gap penalties to find optimal alignment. Hybrid based approaches are also used for better alignment. For example, Stampy use a hash table based algorithm with Single Instruction, Multiple Data (SIMD) based alignment for reporting the best matches [11]. Although each of the tools uses different algorithms for mapping, they have similar performances [12]. However, it is important to select mapping tool, based on the characteristics of the organism, which may otherwise affect downstream analysis.

Several benchmarking analysis came into existence to guide users in choosing aligners. Shang

et al., analyzed the change in computation time with respect to genome size [13]. Martin *et al.*, evaluated the performance of aligners on metagenomics data and determined the limitation of the aligners with respect to the genome size and reference organism within a community to identify each genus independently [14]. Hatem *et al.*, analyzed parameters of mapping with respect to genome comprising speed, sensitivity, read length, type, gap like errors from sequencing technology [12]. Several tools are developed to analyze the performance of aligners. Simulation tools like DWGSIM (http://sourceforge.net/projects/dnaa/), ART [15], Wgsim (https://github.com/lh3/wgsim) are used to simulate reads. Ruffalo *et al*., developed a tool for checking the quality of mapping tools using logistic regression model on mapping statistics [16].

Selection of aligners based on genome characteristics is poorly studied. We extend the evaluations of different aligners based on genome characteristics (genome size, tandem repeats) that induce bias in read alignment. Evaluation of different aligners deployed to map diverse or similar genomes considering various criteria are discussed. Our study can guide users in the selection of a suitable aligner based on genome characteristics.

2. Methods

2.1. Selection of genomes and aligners

The genomes of *Escherichia coli, Saccharomyces cerevisiae*, *Saccharomyces paradoxus, Yarrowia lipolytica and Plasmodium falciparum* with different complexities and distinct characteristics (Table.1) were considered. *E. coli* is a GC-rich gamma-proteobacterium with a single chromosome [17], whereas *P. falciparum* is AT-rich (GC-poor) and has 14 chromosomes [18]. *S. cerevisiae* and *S. paradoxus* [19] are phylogenetically closer and have similar genome size and GC%; whereas *Y. lipolytica* is distantly related and has a larger genome size with a high GC%. Moreover, *Y. lipolytica* [10] has similar genome size as *P.*

falciparum with different chromosome numbers and GC% [20-22]. These five different genomes were considered to elucidate the mapping bias in relation to their genome characteristics such as genome size, tandem repeats, etc.

……..Table.1…….

Aligners were selected based on different algorithms. Some of the aligners included in this study use, indexing based (Bowtie2, BWA), hashing based (NovoAlign, Smalt) and hybrid approach (Stampy). The salient features of these aligners are listed in Table 2.

……..Table.2…….

2.2. Data simulation and processing

The sequencers are rapidly evolving to produce maximum read size so as to reduce mapping bias. Recently, the maximum read-length that could be generated by illumina sequencer has been increased to 300bp (http://systems.illumina.com/systems/sequencing.html). PacBio (http://www.pacificbiosciences.com/) and Nanopore (https://www.nanoporetech.com/) sequencing produces reads of 3kb and 10kb respectively. We generated synthetic paired-end (PE) reads and simulated for illumina (http://systems.illumina.com/systems/sequencing.html) like read sizes of 36bp, 50bp, 72bp, 100bp, 125bp, 150bp, 200bp, 250bp and 300bp for the selected genomes using the read simulator DWGSIM (http://sourceforge.net/projects/dnaa/).

2.3. Estimation of aligner's sensitivity

Efficiency and sensitivity of the aligners in response to read mapping could be affected by the characteristics of genome. We used DWGSIM (dwgsim_eval) to estimate the sensitivity of

these aligners for the simulated reads from the selected genome.

We defined sensitivity (S) as the ratio of reads mapped correctly to the reads mapped incorrectly at a particular threshold, eq. (I)

$$S = \frac{\text{Number of reads mapped correctly}}{\text{Number of reads mapped incorrectly}} \qquad \text{eq. (I)}$$

Paired-end reads are generated by sequencing from both ends of the fragment. An inner distance or insert size is maintained between these ends of the fragment. If the read pairs are aligned with a distance, equal to the inner distance; it is said to be paired properly and the inner distance could be altered to get exact alignment [23]. Some aligners constrain the insert size for maintaining the inner distance. Samtools [24] were employed for post alignment measures, such as percentage of properly paired reads.

2.4. Repeats identification and incorrect alignment

Repetitive DNA or repeats in the genome associated with read alignment creates a technical setback. Repeats are of different shapes and sizes such as interspersed repeats, tandem repeats or nested repeats. Repeats create ambiguity in alignment and in genome assembly [5].
In order to estimate the repeats/homologs within the genome, each genome was divided into sequences of a 100kbp window. Each of these sequences were blast against the whole genome. The E-value and Identity score cutoff were recorded. The sequence that has identical match (100 %) with the target within the same window were discarded. Total repeats in *E. coli, S. cerevisiae, S. paradoxus, Y. lipolytica* and *P. falciparum* were counted as 29,567; 9,702, 9,894; 16,455 and 227,268 respectively; with 80 % identity score. Based on the repeat counts these genomes were segregated as high, intermediate and low repeats genomes.

Tandem repeats are concordant repeats that are predominant in genome sequences. Simple

tandem repeats (STR) are identical sequences that are distributed throughout the genome and the prevalence of these repeats affects the performance of the aligners [25, 26].

In order to investigate the effect of tandem repeats and the incorrect alignment from different aligners, we estimated the correlation coefficient between percentage of tandem repeats and percentage of incorrect alignment within a specific 100kbp window of the genomes using the 5 selected aligners. The tandem repeats were identified using etandem from the emboss package (http://emboss.sourceforge.net/apps/etandem.html) with the parameters set as 80% identity threshold [27].

……..Figure.1……..

To determine the incorrectly mapped reads, a strategy was adapted [16]; taking into account the mapped position of the read with respect to the actual position of read in the chromosome, along with the percentage of match or mismatches. According to this strategy, if the mapped position of the read and actual position of read in the chromosome are the same then the mapping was considered to be correct; irrespective of mismatches in the genome or read (Figure.1. panel a, b, c). With 100% match, the mapping is considered to be correct even if mapped and actual position of read in the chromosome were different (Figure.1. panel d); taking into account the aligners performance in experimental data wherein there is no clue of actual position of reads in the chromosome. With a mutation the mapped positions of the reads could be different from the actual position of the chromosome; and there is a possibility to align incorrectly as shown in (Figure.1. panel 1e, 1f).

2.5.Estimation of computational time for alignment

We have deployed a large-memory server with Intel® Core™ (i7-4770K CPU @ 3.50GHz × 8) processor and a maximum memory of 32 GB of RAM with Gallium 0.4 on NVE7 graphics and Ubuntu 16.04.1 LTS OS for alignment jobs. We were aware that some of these aligners could use multi-threading, in order to avoid any competition between the threads, these aligners were subjected to single-thread mode and computational time of alignments were estimated.

3. Results

3.1. Sensitivity of mapping

About 4 million reads (2 million read pairs) of different size (36bp, 50bp, 72bp, 100bp, 125bp, 150bp, 200bp, 250bp and 300bp were reported and mapped back to the reference genome. This step was repeated with all the selected aligners. Some of the reads were mapped and others were unmapped at a particular threshold. The correctly mapped reads were separated from incorrectly mapped. The ratio of correctly mapped to the incorrectly mapped reads at a particular threshold determines the sensitivity of the aligners (eq.I). Figure.2 shows incorrectly mapped as a function of correctly mapped reads of various size aligned to 5 different genomes using 5 aligners at a particular threshold. The slope of the curve represents sensitivity of the aligners based on different read size and genome type.

……..Figure.2……..

BWA and Smalt showed similar pattern of sensitivity for all the genomes irrespective of the

read size. Bowtie2 showed similar pattern with Stampy. All the aligners when compared with NovoAlign showed lower sensitivity of mapping to the *E.coli* genome (4.6 Mb) (Figure.2, first panel). However, the sensitivity of mapping of these aligners increased with the read size (> 100bp) . A different pattern in sensitivity was observed on mapping reads to the *P.falciparum* genome (22.8 Mb) (Figure.2, second panel). The sensitivity of read alignment to the *P.falciparum* genome for all the aligners including NovoAlign were increased with read sizes (> 100bp). This suggest lack of aligner sensitivity of read alignment to a complex genome which could be eventually enchanced by using shorter reads. Thus, read size greater than 100bp is recommended for these genomes irrespective of aligner selection.

3.2. Properly paired reads

Reads are properly paired when both the mates of the reads are in a proper orientation and has certain insert size on alignment with the genome, it also determines the mapping certainty. Figure.3 shows the percentage of properly pair reads of various sizes, mapped to 5 different genomes with 5 different aligners. The aligners behaves differently, the percentage of properly paired reads spanned between 88 – 99 %, which is color coded with light yellow and brown respectively. BWA, NovoAlign (except with *S. cerevisiae* and *P. falciparum* genomes) and Stampy showed high percentage of properly paired reads (Figure.3). Bowtie2 and Smalt showed lower percentage which substantially decrease with short reads (36bp, 50bp) irrespective of the genome type.

…….. Figure.3 ……..

3.3. Effect of repeats and incorrect alignment

We performed a correlation analysis of percentage of incorrectly mapped reads and percentage of tandem repeats found within a 100kbp window of the genomes of 5 different organisms using the selected aligners.

The *E.coli* genome (low repeats), showed low correlation between repeats and incorrectly mapped reads, irrespective of read size and aligner type. BWA, Stampy had similar value of correlation coefficients irrespective of read size. Percentage of incorrectly mapped reads remained independent of tandem repeats with long reads (100bp, 200bp) when mapped with NovoAlign.

For the *S.paradoxus* genome (intermediate repeats) no correlation was observed between the percentage of tandem repeats and percentage of incorrectly mapped reads when mapped with different aligners. A small increase in correlation coefficient was observed with the *P. falciparum* (high repeats) genome. This suggests lack of specificity of the aligners to deal with the tandem repeats.

3.4.Time comparisons of Aligners

The computational time of read alignments of different aligners were directly proportional to the size of the reference genome. This could be seen for *E. coli* and *P. falciparum* genome (Figure.4, panel first and second) respectively. Apart from the genome size the alignment time also depend on the read size; longer read (> 200bp) took more time irrespective of the reference genome type. A comparative analysis of alignment time showed BWA was significantly faster followed by Bowtie2 and Smalt. NovoAlign and Stampy took longer alignment time (~400 minutes for longer reads) (Figure.4).

…….. Figure.4 ……..

4. Discussions

In NGS analysis alignment of reads to a reference genome is the fundamental step, and selection of the right aligner is therefore important. Aligners are based on different algorithms and they have different approaches to read alignment. However, each genome has unique characteristics (Table.1), repeats in the genome complicate the alignment process [5]. Additionally, reads with relatively shorter size, presence of mismatches and indels pose several challenges in alignment. Performance and specificity of the aligners depends on genome characteristics and we therefore evaluated different aligners and their mapping bias based on several criteria's such as read size, properly paired, tandem repeats and incorrectly mapped reads.

Sensitivity of alignment was affected by genome size, read size and distribution of tandem repeats. Long reads decreased the number of incorrect alignment and increased the sensitivity of alignment for all the aligners. Random alignment of shorter reads (36bp, 50bp or 75bp) in the genome without *a-priori* knowledge of genome location was obeserved. Among different selected aligners, NovoAlign was found to have higher sensitivity with short and long reads. BWA, Smalt showed similar pattern of sensitivity. Stampy behaves differently and showed similar sensivity pattern with Bowtie2 (expect for *P. falciparum*). Overall, low sensitivity was found with short reads (36bp, 50bp), which was true for all the aligners.

The preponderance of STR, tricks the aligners that hinders effective mapping of reads to the

reference genome. *Hongsoek et al*. discussed aligners that assign high-quality scores to the incorrectly aligned reads, when reported with two or more tandem repeats loci with same motif but different repeat lengths [28]. High coverage due to repeats has been known [29] and tandem repeats in the genome plays a critical role in assessing the quality of the aligners. Correlation analysis suggested, low specificity of the aligner's to classify incorrectly mapped reads associated with tandem repeats within a specified genome window.

Comparative analysis of alignment time showed BWA was significantly faster followed by Bowtie2 and Smalt. NovoAlign and Stampy were comparatively slower and took about ~400 minutes for alignment of longer reads to *P. falciparum* genome.

Various features of the genome or reads either independently or in combination should be preferentially weighed by the aligners for an effective alignment. Considering all these factors the five different aligners were evaluated, assessed and benchmark with a scoring systems depicted in Table.3.

Alignment with short reads still remains an active challenge. Long reads enhance the sensitivity of mapping to a certain extent. Bowtie2 and Smalt showed improper paring of the read mates. The aligners showed a low specificity on aligning tandem repeats. All these factors together are essential to enhance the effectiveness of mapping. The end user therefore has to choose aligner based on different factors including genome features. Improved aligners have to be designed that would take into account these features and hereby enhances the accuracy of read alignment.

Competing interests

The authors declare no competing financial interests.

Authors' contributions

STK designed the study, carried out analysis and participated in manuscript preparation. PS designed the study, performed statistical analysis and participated in manuscript preparation. IN conceived and supervised the study edited the manuscript. All authors have read and approved the manuscript.

Conflict of interest

The authors declares that there is no conflict of interest.

Acknowledgements

We would like to thank Prof. Jens Nielsen for valuable suggestions and useful scientific discussions. This work was financially supported by the Knut and Alice Wallenberg Foundation, BILS and Vetenskapsrådet.

References

1. Nowrousian M: **Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems**. *Eukaryot Cell* 2010, **9**(9):1300-1310.
2. Lee C-Y, Chiu Y-C, Wang L-B, Kuo Y-L, Chuang EY, Lai L-C, Tsai M-H: **Common applications of next-generation sequencing technologies in genomic research**. *Translational Cancer Research* 2013, **2**(1):33-45.
3. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nat Rev Genet* 2011, **12**(6):443-451.
4. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
5. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions**. *Nat Rev Genet* 2012, **13**(1):36-46.
6. Lindner R, Friedel CC: **A comprehensive evaluation of alignment algorithms in the context of RNA-seq**. *PLoS One* 2012, **7**(12):e52403.
7. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

8. Marco-Sola S, Sammeth M, Guigo R, Ribeca P: **The GEM mapper: fast, accurate and versatile alignment by filtration**. *Nat Methods* 2012, **9**(12):1185-1188.
9. Kent WJ: **BLAT—the BLAST-like alignment tool**. *Genome research* 2002, **12**(4):656-664.
10. Dujon B: **Yeast evolutionary genomics**. *Nat Rev Genet* 2010, **11**(7):512-524.
11. Lunter G, Goodson M: **Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads**. *Genome Res* 2011, **21**(6):936-939.
12. Hatem A, Bozdag D, Toland AE, Catalyurek UV: **Benchmarking short sequence mapping tools**. *BMC Bioinformatics* 2013, **14**:184.
13. 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**:309650.
14. Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, Orvis J, Sodergren E, Wang Z, Weinstock GM *et al*: **Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities**. *PLoS One* 2012, **7**(6):e36427.
15. Huang W, Li L, Myers JR, Marth GT: **ART: a next-generation sequencing read simulator**. *Bioinformatics* 2012, **28**(4):593-594.
16. Ruffalo M, Koyuturk M, Ray S, LaFramboise T: **Accurate estimation of short read mapping quality for next-generation genome sequencing**. *Bioinformatics* 2012, **28**(18):i349-i355.
17. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF *et al*: **The complete genome sequence of Escherichia coli K-12**. *Science* 1997, **277**(5331):1453-1462.
18. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al*: **Genome sequence of the human malaria parasite Plasmodium falciparum**. *Nature* 2002, **419**(6906):498-511.
19. Sherman D, Durrens P, Iragne F, Beyne E, Nikolski M, Souciet J-L: **Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts**. *Nucleic acids research* 2006, **34**(suppl 1):D432-D435.
20. Sen P, Vial HJ, Radulescu O: **Mathematical Modeling and Omic Data Integration to Understand Dynamic Adaptation of Apicomplexan Parasites and Identify Pharmaceutical Targets**. *Comprehensive Analysis of Parasite Biology: From Metabolism to Drug Discovery* 2016:457.
21. Sen P, Vial HJ, Radulescu O: **Kinetic modelling of phospholipid synthesis in Plasmodium knowlesi unravels crucial steps and relative importance of multiple pathways**. *BMC systems biology* 2013, **7**(1):123.
22. Sen P: **Integrated modelling of lipid metabolism in Plasmodium, the causative parasite of malaria**. Université Montpellier II-Sciences et Techniques du Languedoc; 2013.
23. Ratnakumar A, McWilliam S, Barris W, Dalrymple BP: **Using paired-end sequences to optimise parameters for alignment of sequence reads against related genomes**. *BMC genomics* 2010, **11**(1):458.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
25. Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, Balasubramanian S,

Boden M: **Inferring short tandem repeat variation from paired-end short reads**. *Nucleic Acids Res* 2014, **42**(3):e16.

26. Ummat A, Bashir A: **Resolving complex tandem repeats with long reads**. *Bioinformatics* 2014, **30**(24):3491-3498.
27. Levdansky E, Romano J, Shadkchan Y, Sharon H, Verstrepen KJ, Fink GR, Osherov N: **Coding tandem repeats generate diversity in Aspergillus fumigatus genes**. *Eukaryot Cell* 2007, **6**(8):1380-1391.
28. Tae H, McMahon KW, Settlage RE, Bavarva JH, Garner HR: **ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats**. *Bioinformatics* 2013, **29**(14):1734-1741.
29. Misawa K: **RF: a method for filtering short reads with tandem repeats for genome mapping**. *Genomics* 2013, **102**(1):35-37.

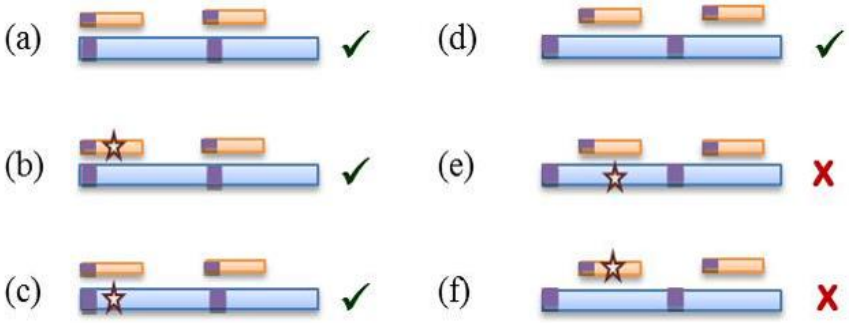

Figure.1. Strategy to determine incorrect reads. Blue bars indicate the genome, orange bars indicate the reads and purple blocks indicate the position of reads mapped and actual position of read in the chromosome, star indicate the presence of mismatches. Tick marks are considered as correct alignments and cross marks as incorrect alignments.

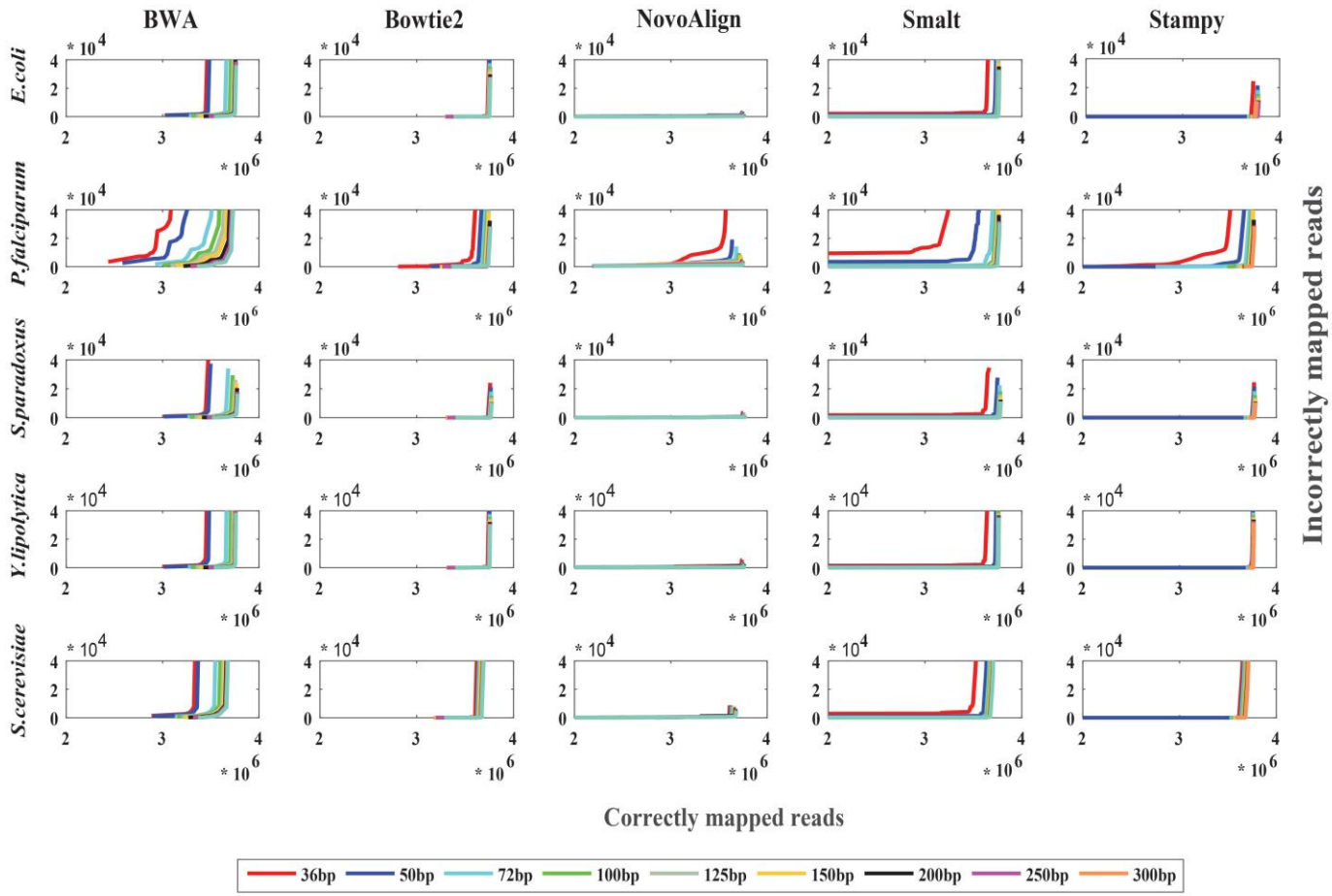

Figure.2. Sensitivity curves that represents incorrectly mapped vs. correctly mapped reads of variable size illustrated with different colors, mapped to 5 different genomes (rows) using 5 different aligners (columns).

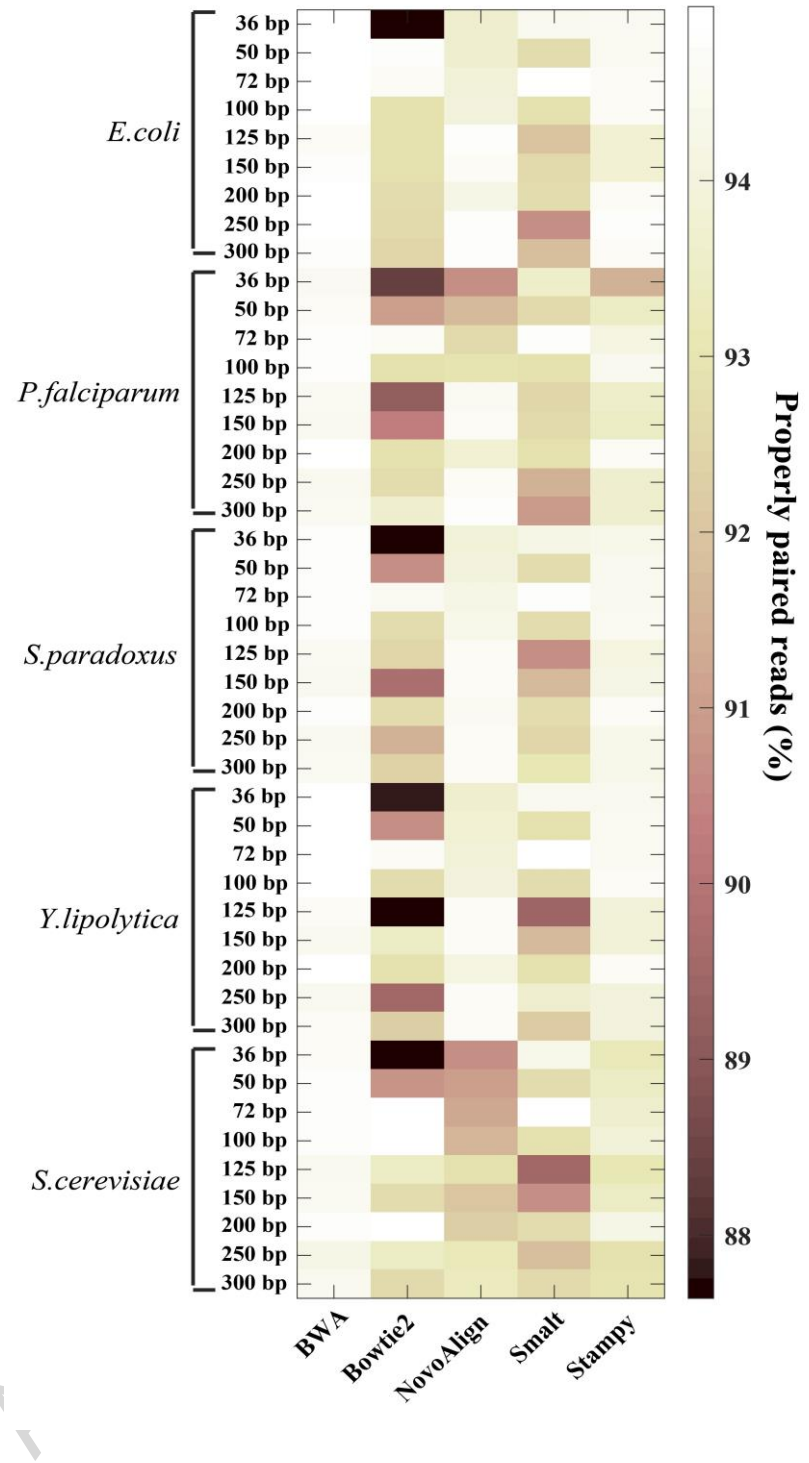

Figure.3. Each line in the heatmap represents percentage of properly paired reads of a particular size aligned to genome with five different aligners.

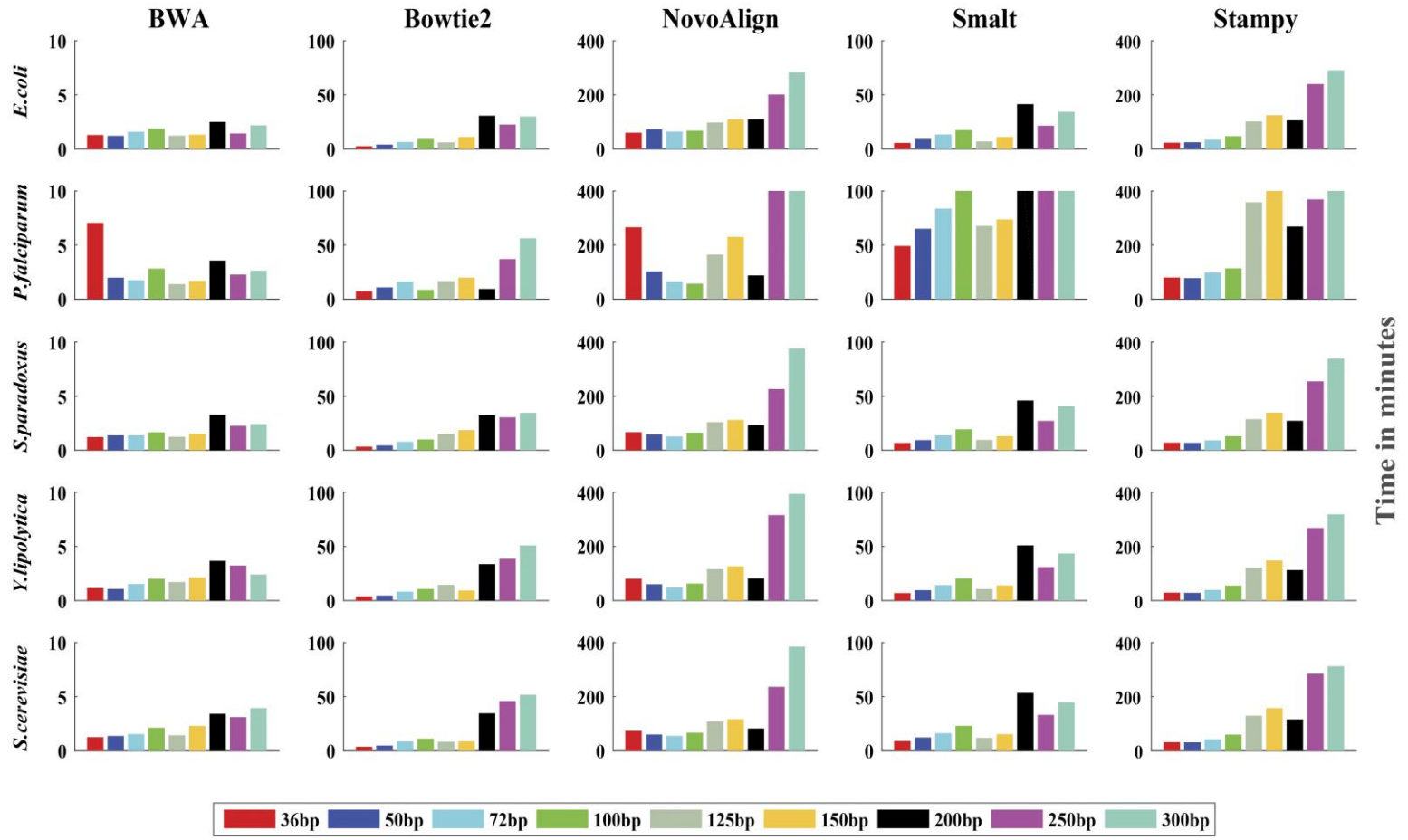

Figure.4. Computational Time (minutes) estimated for reads of variable sizes mapped to reference genomes (rows) with five different aligners (columns).

Sl.No	Tools	Indexing/ Hashing	Algorithms used	Salient features		
		Organism	**Chromosome *numbers***	**GC %**	**Genome size**	**Average Tandem repeats** (in each Chromosome)
		E. coli K12 MG155	1	50.79	4.6 Mb	0.35%
		P. falciparum 3D7	16	20	22.8 Mb	32%
		Yarrowia Lipolytica CLIB122	6	49.08	20.5 Mb	0.5 %
		S.paradoxus CBS432	16	38.06	11.7 Mb	0.46%
		S. cerevisiae s288c	16	38.03	12 Mb	0.61%

Table.1. Genome characteristics of the organisms

1	BWA	FM-index	Burrows–Wheeler transform and Smith-Waterman method Prefix/Suffix Matching Algorithms	BWA-backtrack (for shorter reads) BWA-SW, BWA-MEM (for longer reads). Generally used for mapping less divergent sequence. BWA-SW allows gaps in seeds.
2	Bowtie2	FM-index	Modified Ferragina and Manzini matching algorithm BWT-indexing Prefix/Suffix Matching Algorithms, quality aware backtracking	Allows gapped alignment and compared with Bowtie 1, its sensitivity is high for reads greater than 50bp.
3	NovoAlign	Hashing	Needleman-Wunsch algorithm, Iterative Search alignment scoring	Alignment quality scores using posterior alignment probability. Reports multiple alignments per read.
4	Smalt	Hashing	Smith Waterman algorithm and short word hashing	Provide tuning parameters indexing word length and step size to improve sensitivity and accuracy.
5	Stampy	Hybrid mapping algorithm	Fast hashing algorithm with statistical approach	Maps read containing sequence variation. Can map highly divergent species.

Table.2. List of aligners selected for the benchmarking study.

	Sensitivity		Properly paired		Computational time		Tandem repeats	
	(36, 50, 72 bp)	(100, 125, 150 200, 250, 300 bp)	(36,50, 72 bp)	(100, 125, 150 200, 250, 300 bp)	(36,50, 72 bp)	(100, 125, 150 200, 250, 300 bp)	Low	High
BWA	+	+++	++	+++	+++	+++	++	+
Bowtie2	+	+++	+	+	++	++	++	+
NovoAlign	+++	+++	++	+++	+	+	++	+
Smalt	+	+++	+	+	++	++	++	+
Stampy	++	+++	++	+++	+	+	++	+

Table.3. Table depicts the overall scoring of the aligners based on various evaluation criteria considered in this study; +++ denotes high score, ++ denotes intermediate score, + denotes low score.

Research highlight: Evaluation and assessment of read-mapping by multiple Next-generation sequencing aligners based on genome-wide characteristics

The five widely used aligners BWA, Bowtie2, NovoAlign, Smalt and Stampy were evaluate their performance on different five microbial genomes, which have diverse genome characteristics. The result form the benchmarking was reported as a guideline for the end user to choose an appropriate aligner enhancing the accuracy of read mapping.