Chiara Becht, Jonas Schmidt, Frithjof Blessing and Folker Wenzel*

Comparative analysis of alignment tools for application on Nanopore sequencing data

Abstract:

INTRODUCTION: Long-read sequencing techniques such as Oxford Nanopore sequencing, are representing a promising novel approach in molecular-biological methodology, enabling potential facilitation in mapping and de novo assembly. In comparison to conventional sequencing methods, novel alignment tools are mandated to compensate differing data structures (especially high error rate) to achieve acceptably accurate analysis results. **METHODS:** In this study, benchmarking for long read aligners BLASR, GraphMap, LAST, minimap2, NGMLR and the short-read aligner BWA MEM on three experimental datasets was conducted. Obtained alignment results were compared for various quality and performance criteria, such as match rate, mismatch rate, error rate, working memory usage and computational time. **RESULTS:** The comparison yielded differences in alignment quality and performance of tools under test. Tool LAST showed the largest differences among all tools. Minimap2 achieved constant quality with good performance. BLASR, GraphMap, BWA MEM and NGMLR showed slight differences only. **CONCLUSION:** Differences among the tools could be reasoned with dataset characteristics and algorithm approaches of individual tools. All tools except BLASR seem applicable for Nanopore sequencing data. Therefore, selection of the tool should be done under consideration of the experimental design and the further downstream analysis.

Keywords: Alignment tools, Nanopore sequencing, Benchmarking, comparative analysis

<https://doi.org/10.1515/cdbme-2021-2212>

*Corresponding author: **Folker Wenzel:** Faculty of Medical and Life Sciences, Hochschule Furtwangen, Schwenningen, Germany, wfo@hs-furtwangen.de

Chiara Becht: Faculty of Medical and Life Sciences, Hochschule Furtwangen, Schwenningen, Germany

Jonas Schmidt, Frithjof Blessing: MVZ Laborärzte Singen, Singen, Germany

1 Introduction

Sequencing of nucleic acids has advanced to a powerful instrument in molecular biological research and medical diagnostics over time. Developments over three generations evolved in various sequencing techniques, all with the goal of optimizing the core-parameters including sequencing speed and parallelisation, costs, maximum read length and overall error rate [1]. One of the latest technologies is Nanopore sequencing which was commercialized by Oxford Nanopore Technologies (ONT). By applying biological nanopores to translate the sequence information into specific electrical current patterns, this approach offers ultra-long reads accompanied by high sequencing speed and low cost [1]. So far, the most important disadvantage of this technology is the increased error rate compared to other sequencing technologies. However, due to continuous development and improvements of the technology until today the difference has become comparably small [1]. The unique error profile makes the analysis of Nanopore sequencing data challenging and requires specialized algorithms. Handling of errors during different data analysis steps is crucial for the reliability of the result and the final interpretation [2]. One fundamental step in many data analysis pipelines is sequence alignment, which summarizes the process of referring the reads back to a reference sequence [3]. Regarding Nanopore sequencing data, there is a wide variety of specialised alignment tools freely available, which can be used for sequence alignment. Therefore, it is important to compare their performance prior to use and select the best suitable tool for a specific project. This has already been done by others using artificially created dataset [4]. In this work, we focus on the comparison of six state of the art alignment tools with regard to their performance and alignment quality using three different datasets from biological Nanopore sequencing experiments. Further, to account for the flexible use of Nanopore sequencers the comparison was performed using only conventional hardware to test the applicability under standard laboratory conditions.

2 Material and Methods

2.1 Datasets

The test datasets used for comparison of the different alignment tools originate from three different sequencing experiments including I) lambda phage whole genome sequencing, II) amplicon sequencing of the human pyrin innate immunity regulator (MEFV) gene and III) severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) whole genome sequencing (Table 1). All of them were produced during sequencing experiments on a MinION sequencing device using R9.4 flow cells in combination with ligation sequencing following the manufacturers protocol (ONT, Oxford, England). They mainly differ in quality, size and read length distribution.

Prior to the use in this study all datasets were completely anonymized to remove any traceability to a specific sample.

The datasets were prepared for the comparison of the alignment tools in FASTQ file format after base calling raw sequencing data using a high accuracy base calling model of the base caller Guppy [v3.1.5] (ONT). In order to remove chimeric reads, prior to further analysis length filtering based on the expected read length was performed for the datasets originating from amplicon sequencing experiments.

Reference sequences for the alignments were downloaded from the NCBI reference sequence database (SARS-CoV-2 genome: NC_045512.2, human chromosome 16: NC_000016.10, lambda phage genome: NC_001416.1).

Table 1: Characteristics of the three experimental datasets used for the comparison of the alignment tools. All datasets originate from Nanopore sequencing experiments performed on a MinION sequencing device.

Dataset	Target	DNA input	Median Q Score	Total read count	Median read length [bases]
Lambda	Whole Genome	gDNA	13.1	134534	9123
MEFV	MEFV Gene	Amplicons	11.4	154664	463
SARS-CoV-2	Whole Genome	Amplicons	13.2	86418	381

2.2 Computational setup

The comparison of the different alignment tools was conducted on a conventional notebook (Intel Core i7-8565U,

16 GB Random-Access Memory (RAM), 512 GB Solid State Drive (SSD)) running Ubuntu (Ubuntu 18.04.5 LTS) and Microsoft Windows (Windows 10pro version 2004). Base calling was performed on a separate high-performance workstation (Intel Core I7 7700K, 64 GB RAM, Nvidia Geforce RTX2080 TI, 1 TB SSD).

2.3 Comparative analysis

Six different alignment tools were incorporated into the comparison including the dedicated long read alignment tools BLASR [v5.3.3] (github.com/PacificBiosciences/blasr), minimap2 [v2.17-r941] (github.com/lh3/minimap2), LAST [v1060] (github.com/mcfrith/last-genome-alignments), GraphMap [v0.5.2] (github.com/isovic/graphmap) and NGMLR [v0.2.7] (github.com/philres/ngmlr) as well as the state-of-the art short-read alignment tool BWA MEM [v0.7.17-r1188] (github.com/lh3/bwa). For the comparative analysis the tools were used with the default configuration for Nanopore sequencing data as stated in the respective usage manual. In order to compare the performance of the different tools all three datasets were aligned to their respective reference sequence by using each alignment tool and the output was stored in Binary Alignment Map (BAM) files or if not supported by the tool in tab delimited files.

Performance measures including computational time (measured as Central Processing Unit (CPU) time) and working memory consumption (measured as peak RAM usage) of each individual run were recorded by using the time command in Ubuntu. The measurements were performed in single-thread mode and multi-thread mode, when supported by the alignment tool.

The frequencies of matches (identical nucleotides between query and reference), mismatches (differing nucleotides between query and reference), deletions (missing nucleotides in query compared to reference) and insertions (additional nucleotides in query compared to reference) were calculated for each alignment from the alignment output files by using custom R scripts. They were used to calculate quality indicators including match rate (see eq 1), mismatch rate (see eq 2) and error rate (see eq 3).

Median match, error and mismatch rates were statistically compared by applying a generalized linear model (GLM) assuming a Poisson distribution on the underlying count data and using the alignment length as an offset variable.

All data analysis and visualization were done using R [v3.6.3] (R Foundation for Statistical Computing, Vienna, Austria).

$$\text{Match rate} = \frac{\sum \text{matches}}{\sum \text{matches} + \text{mismatches} + \text{deletions} + \text{insertions}} \quad (1)$$

$$\text{Mismatch rate} = \frac{\sum \text{mismatches}}{\sum \text{matches} + \sum \text{mismatches} + \sum \text{deletions} + \sum \text{insertions}} \quad (2)$$

$$\text{Error rate} = \frac{\sum \text{mismatches} + \sum \text{insertions} + \sum \text{deletions}}{\sum \text{matches} + \sum \text{mismatches} + \sum \text{deletions} + \sum \text{insertions}} \quad (3)$$

3 Results

Computational performance of the different alignment tools was assessed by recording CPU time and peak RAM usage (Figure 1). Summarized, all three datasets were successfully processed by each tool on a conventional notebook. The Lambda dataset required the longest median processing time (1913.49 seconds (Lambda) compared to 184.34 seconds (MEFV) and 15.27 seconds (SARS-CoV-2)) and showed the highest median peak RAM usage compared to the other datasets (1760.97 MB (Lambda) compared to 877.68 MB (MEFV) and 75.10 MB (SARS-CoV-2)). Multithreading decreased the CPU time with an increase of the peak RAM usage. The comparison of the different tools showed a superior performance of minimap2 regarding speed and memory consumption on all three datasets.

The performance of the other alignment tools varied between the different datasets. Overall, BWA MEM and LAST showed intermediate CPU time and peak RAM usage. BLASR, GraphMap and NGMLR showed the highest CPU times with varying peak RAM consumption. Especially NGMLR had high peak RAM values.

Figure 1: A) CPU time of the different alignment tools running on 1, 2, 4 and 8 threads. B) RAM usage of the different alignment tools when running on 1, 2, 4 and 8 threads. Multithreading was not supported by minimap2.

To compare the alignments produced by the different tools for the three datasets quality measures including match rate, mismatch rate and error rate of the alignments were extracted from the output files.

BLASR was excluded from the calculations of the MEFV dataset due to insufficient position information. Therefore, an analysis of the MEFV amplicon sequencing data under the

Figure 2: Distribution of A) match rate, B) mismatch rate and C) error rate for the application of the tested alignment tools on all three datasets. BLASR was excluded from the calculations of the MEFV dataset due to insufficient position information, which made the comparative analysis of MEFV exon amplicon sequencing data impossible for this tool.

same conditions as for the other tools was not possible.

All tools except LAST showed a similar distribution of the three rates over all alignments generated per dataset (Figure 2). The achieved match rates ranged mainly between 85 % and 100 % (Figure 2A). The GLM showed no significant influence of the tool choice on the median match rate for all of the datasets. Mismatch rates were mainly observed in the range of 0 % to 5 % for all three datasets (Figure 2B). By applying the GLM a highly significant positive influence (Estimate [E] = 1.145, Standard Error [SE] = 0.155, P < 0.001) of the tool LAST on the median mismatch rate was observed for the Lambda dataset. Additionally, a significant positive influence (E = 0.353, SE = 0.130, P = 0.007) of the tool BWA MEM was observed. For the other two datasets the GLM showed no significant influence of the tool on the median mismatch rate. The error rate which includes mismatches, insertions and deletions ranged mainly between 0 % and 12 % (Figure 2C). Again, tool LAST showed a highly significant positive influence (E = 0.405, SE = 0.100, P < 0.001) on the median error rate for the Lambda dataset. For the other two datasets no significant influences of the tools were observed.

4 Discussion

To evaluate the performance and quality of modern alignment tools for Nanopore sequencing data we compared six long-read alignment tools by using three experimental datasets. The three different datasets were selected to account for varying data structure regarding read length distribution, read count and overall read quality. Computational performance of the different tools was assessed by recording CPU time and peak RAM usage on a conventional notebook during the alignment of the test datasets to the respective reference sequence. Remarkably, by using only standard hardware none of the tools failed to process the alignments. This is a promising result since one big advantage of Nanopore sequencing is the portability of some sequencing devices. To support this feature, it is important that the tools used for data analysis can be applied on standard hardware without the need of high-performance workstations. As expected, the use of multithreading led to reduced runtime with an increased peak RAM consumption. This is important since further process parallelization can seriously speed up the whole data analysis workflow. Major differences in CPU time and peak RAM usage were observed between the different tools. Even without the option of multithreading minimap2 was the fastest tool on all three datasets with an intermediate peak RAM consumption. BLASR, GraphMap and NGMLR showed comparably long run time. This let them appear to be less efficient for the application on long read datasets. In addition, NGMLR showed a high peak RAM usage which might be a bottleneck for larger datasets on less powerful systems. In general, differences in performance between the different tools can be explained by different algorithmic approaches. However, since all tools proved to be functional even with standard hardware, performance should play only a minor role in the selection of a specific alignment tool for a certain experiment. In order to compare the alignments produced by the different tools quality measures including match rate, mismatch rate and error rate were calculated from the output files. Overall, the analysis revealed high match rates and low mismatch and errors rates for all datasets when applying BLASR, BWA MEM, minimap2, GraphMap and NGMLR. This indicates that these tools can be applied for the analysis of Nanopore sequencing data. LAST showed a significantly higher mismatch rate and error rate for the Lambda dataset compared to the other tools. Similar observations were also made by others [5, 6]. Since the Lambda genome is comparably small, long reads should be easy to map. Considering the high match rates and low mismatch/error rates achieved by the other tools, the difference observed for LAST seems to originate from the tool itself [6]. Possible

explanations are the use of scoring algorithm and the formation of local alignments [6]. Although BLASR showed an acceptable performance for the Lambda and the SARS-CoV-2 dataset, this tool can only hardly be implemented in data analysis pipelines for Nanopore sequencing data since output in BAM format is not supported for this kind of data. Benchmarking of alignment tools is frequently done by using synthetic datasets [4]. In this study, we used experimental datasets derived from MinION sequencing runs to evaluate the tools under real experimental conditions. Although this approach helps to evaluate and compare the characteristics of an alignment tool, it has the limitation that true known reference values are missing. Therefore, it is not possible to assess quality criteria like precision and recall of the alignment tools [4]. For extended comparison studies further quality measures including clipping rate, read splitting and the mapping position relative to the reference genome can be included. Summarized, in this work by using experimental datasets we show that all evaluated alignment tools except BLASR are suitable for the analysis of Nanopore sequencing data and the integration in a data analysis pipeline. Especially minimap2 showed a superior performance. Differences among the tools could be reasoned with dataset characteristics and algorithm approaches of individual tools. Therefore, selection of the tool should be done under consideration of the experimental design and further downstream analysis.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: None declared. Informed consent: No individuals were included in this study. Ethical approval: No research related to human use was conducted. The datasets were completely anonymised prior to the use in this study.

References

- [1] Schmidt J, Blessing F, Fimpler L, Wenzel F. Nanopore Sequencing in a Clinical Routine Laboratory: Challenges and Opportunities. *Clin Lab* 2020;66(6).
- [2] Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. *Nat Biotechnol.* 2016;34:518–24.
- [3] Pevsner J. *Bioinformatics and functional genomics*. 3rd ed. Chichester: Wiley Blackwell; 2015.
- [4] Pavlovikj N, Moriyama EN, Deogun JS. Comparative analysis of alignment tools for nanopore reads. *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*. 2017:169–74.
- [5] Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* 2016;7:1–11.
- [6] Li H. A few comments on GraphMap. 2015. <https://lh3.github.io/2015/07/30/a-few-comments-on-graphmap>. Accessed 21 Jan 2021.