CHALLENGES AND PROMISES OF NANOPORE SEQUENCING BASED HLA GENOTYPING

Author: Réka Nagy[1]

1. Omixon Biocomputing Kft, Budapest, Hungary

Introduction

The introduction of Nanopore sequencing opened a whole new world for bacteria and eukaryotic genome sequencing with promises of speed, significant on-demand usability and low cost in comparison to well known short-read sequencing methods. ONT (Oxford Nanopore Technologies) announced an EAP for the R10 flow cell chemistry in March 2019, which is expected to become the successor of R9 chemistry, introduced in 2016. The different flow cell versions describe the nanopores used for long-read sequencing. The widespread R9 nanopores are derived from a mutant of a CsgG lipoprotein from E. Coli, which provides transport for polypeptides across the bacterial membrane. This structure was engineered to allow the transport of DNA instead of peptides. [1] In comparison to R9, R10 pores have a longer barrel and a dual reader head (Fig. 1) that aim to specifically improve the resolution of homopolymer regions and consensus accuracy. [2] As the ssDNA molecule traverses through the membrane, the changes in the ionic current are measured and the traversed nucleotides are identified based on this measured profile. There are 512 nanopore channels on the MinION sequencing platform, which function independently and are controlled by a standard laptop computer (Fig. 2).

The utilization of Nanopore sequencing in HLA typing has great potential to reduce resource costs in a lab but more importantly, the long reads could resolve common challenges of short-read sequencing. More and more studies are demonstrating promising results showing that Nanopore sequencing can overcome the obstacles set by this challenging gene family.

The goal of the work presented here is not only to compare the performance of flow cells version R9.4.1 and R10, but also to assess them in comparison to short-read data (MiSeq, Illumina).

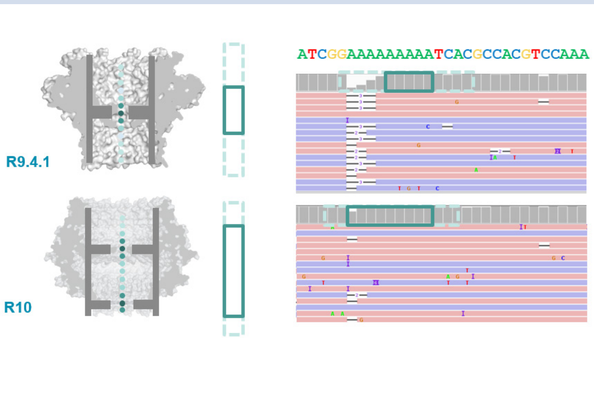

FIGURE 1 - Comparison of different nanopore structures and their performance on homopolymer regions. [2]

FIGURE 2 - The MinION sequencing platform by ONT weights 103 g with the loaded flow cell and only 105 mm long.

Methods

A 24-sample, 7-locus library (HLA-A, HLA-B, HLA-C, HLA-DPB1, HLA-DQA1, HLA-DQB1 and HLA-DRB1) was prepared using the Holotype HLA v3 24/7 kit configuration and sequenced on Illumina's MiSeq using the 300-cycle chemistry. Three of these samples that are well characterized cell lines (JVM, MOU, PGF) with known HLA genotypes were selected for Nanopore sequencing. The amplicons of these 3 samples were prepared were prepared for Nanopore sequencing using the Ligation Sequencing Kit (SQK-LSK109) and the PCR Barcoding Kit (SQK-PBK004). The libraries were sequenced on two MinIONs using R9.4.1 and R10.0 flow cells respectively. The three samples were run in duplicate to assess reproducibility. All the data was basecalled using the high-accuracy basecalling mode of the Guppy algorithm.

Both short- and long-read sequencing data was aligned to the reference sequences with bwa (v0.7.12-r1039). Based on the alignments, consensus sequences were generated with majority rule. The result consensus sequences were then compared to the reference alleles and the concordance was calculated for each of them. The result alignment files were assessed with the BAMStats tool (v1.25) [3]. The short-read data was analyzed with Omixon HLA Twin 3.1.1 CE and IMGT/HLA database v3.31 with default analysis settings to generate consensus sequences. De novo assembled sequences were generated from the Nanopore reads with a combined pipeline using Pomoxis (v0.3.0a1) [4] and Medaka (v0.11.1) [5] (Fig. 3). Finally, these consensus sequences were assessed with Pomoxis and the result error profiles were compared.

The plots summarizing the results were created with the ggplot library of R.

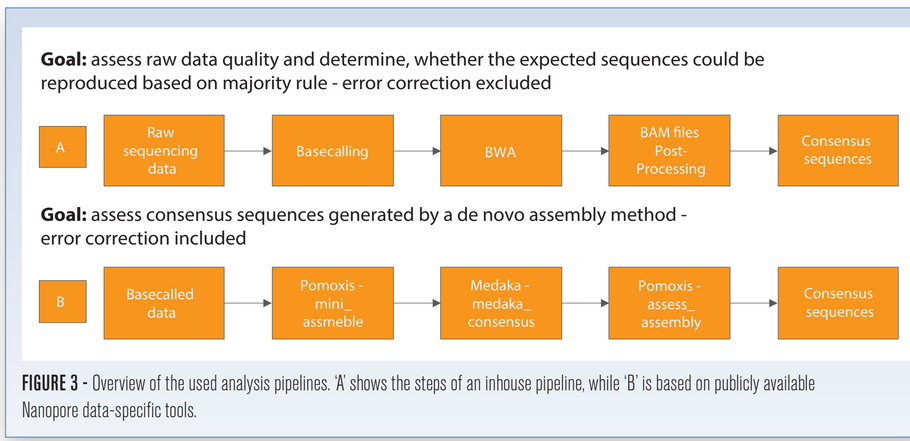

FIGURE 3 - Overview of the used analysis pipelines. 'A' shows the steps of an inhouse pipeline, while 'B' is based on publicly available Nanopore data-specific tools.

Results

The bwa-based pipeline produces consensus correctness values (Table 1), which describe the quality of our data as the consensus bases are determined based on majority rule. With this approach, each position of the result alignment is inspected and the base with the highest read support is called. As this pipeline does not utilize any data correction function or data pre-processing step, it can be used for describing the actual quality and correctness of the raw sequencing data. While alignment files are generated by bwa, metrics such as the average of mapped read lengths, the coverage depth and read counts can also be assessed with the BAMStats tool. Because the full-length reference sequences used for alignment generation, the consensus correctness values were calculated considering only the targeted regions specified by the locus-specific primers. This way alignment artifacts present at the 5' and 3' consensus ends were excluded. The values were determined for each locus of each sample. The combined pipeline of Pomoxis and Medaka can generate and analyze de novo consensus sequences and determine error types. The 3-step pipeline handles both result generation and assessment, so the results can be directly compared between the short- and long-read data. In contrast to the previous pipeline, this workflow iteratively corrects erroneous bases of Nanopore reads. The assess_assembly command of the Pomoxis tool is able to determine consensus accuracy (Table 2) as well as both allele and sample-level error profiles including substitution, insertion and deletion errors (Table 3). In an effort to compare the metrics of the long-read data to the corresponding ones of the short-read ones, we analyzed the consensus sequences generated by HLA Twin with the same command. The different pipelines produced consistent results.

PER EXPERIMENT CONSENSUS CORRECTNESS VALUES					
	Reference	R9.4.1		R10.0	
Locus		Sample 1	Sample 2	Sample 1	Sample 2
HLA-A	99.78%	99.64%	99.87%	99.35%	99.73%
HLA-B	99.83%	99.70%	99.53%	99.55%	99.44%
HLA-C	100.00%	99.28%	99.61%	99.35%	99.21%
HLA-DPB1	99.37%	97.97%	97.89%	97.09%	97.28%
HLA-DQA1	99.87%	99.86%	99.85%	99.82%	99.91%
HLA-DQB1	99.99%	99.63%	99.60%	99.51%	99.38%
HLA-DRB1	99.02%	93.40%	92.61%	92.47%	94.52%
Grand Total	99.71%	98.52%	98.45%	98.22%	98.56%

TABLE 1 - Per experiment consensus correctness values. Consensus correctness is the ratio of correct bases in the consensus sequence and the consensus length

PER EXPERIMENT CONSENSUS ACCURACY VALUES			
Locus	Reference	R9.4.1	R10.0
HLA-A	100.00%	99.88%	99.11%
HLA-B	100.00%	99.64%	98.02%
HLA-C	100.00%	99.66%	94.80%
HLA-DPB1	99.97%	99.41%	99.64%
HLA-DQA1	100.00%	99.75%	99.26%
HLA-DQB1	99.99%	99.56%	99.43%
HLA-DRB1	99.97%	99.82%	99.29%
Grand Total	99.99%	99.69%	98.25%

TABLE 2 - Per experiment consensus accuracy values. Consensus accuracy is calculated by the pomoxis tool, the following way: Consensus accuracy = 100 - 100 * (insertions + deletions + substitutions) / consensus length

AVERAGE OF OBSERVED DELETION RATES			
Samples	Reference	R9.4.1	R10.0
JVM	0.00%	0.04%	0.19%
MOU	0.00%	0.04%	0.42%
PGF	0.01%	0.06%	0.38%
Grand Total	0.00%	0.05%	0.31%

AVERAGE OF OBSERVED INSERTION RATES			
Samples	Reference	R9.4.1	R10.0
JVM	0.02%	0.34%	0.14%
MOU	0.00%	0.14%	0.32%
PGF	0.00%	0.07%	0.66%
Grand Total	0.01%	0.18%	0.34%

AVERAGE OF OBSERVED SUBSTITUTION RATES			
Samples	Reference	R9.4.1	R10.0
JVM	0.01%	0.06%	0.17%
MOU	0.00%	0.02%	1.01%
PGF	0.00%	0.04%	0.96%
Grand Total	0.00%	0.05%	0.64%

TABLE 3 - The following tables show the rate of different error types observed with short- and long-read data.

Discussion

The significantly higher read count observed with the short-read data (Fig. 4) is counterbalanced by the long-reads of Nanopore data (Fig. 5) thus the average coverage depths values differ to a lesser extent (Fig. 6). The read lengths of Nanopore data also correlate well with the amplicon lengths, the reads aligned to Class II genes are almost twice as long as the ones mapped to Class I genes. This means that a single read can span the whole Class II amplicons without any issue. The majority rule based consensus sequences with the highest scores can be seen in case of the short-read data and only a slight difference can be observed between the values belonging to R9.4.1 and R10.0 flow cells. The average correctness values for long-read data reach beyond 98%.

When comparing the results of the combined pipeline of Pomoxis and Medaka a slight difference can be observed between the flow cell versions. The results of R9.4.1 outperform R10.0 for each inspected metric. With R10.0 chemistry, a pore utilization problem is also observed during sequencing. Despite the fact that the same number of total read counts are observed in each sample, the samples with R10.0 chemistry reached the 76 000 target reads 3-4 times slower. With one of R10.0 samples, the sequencing took approximately ~17.5 hours to complete and reach target value.

Conclusion

The work presented here focuses on the comparison of the different flow cell chemistries offered by ONT, as well as the comparison of short-read and long-read sequencing data of the same samples for 7 HLA loci. Comparing the different types of data using alignments to known reference sequences reveals valuable information about the resolution of issues associated with homopolymer regions, which is still a challenge for any sequencing platform. The long reads effectively allow us to straightforwardly phase distant heterozygous variants, thus resolving ambiguities previously limited by the short-read approach. By comprehending our data, assessing error profiles and differentiating random and systematic noise content we can develop proper analysis methods and propagate the potential of Nanopore sequencing in everyday clinical routine.

Although the accuracy of consensus sequences of Nanopore data is still behind the ones of short-read data, the analysis tools show promising results and they are continuously upgraded to achieve even better results. Based on the results, the R10.0 chemistry did not prove to be the potential successor of R9.4.1 due to the lower capture and quality of reads. ONT already introduced the R10.3 flow cell, which replaces R10.0 since mid-December of 2019. The R10.3 promised to have single-molecule accuracy as good as R9.4.1 and can be used with the PromethION sequencing platform in contrast to R10.0 leaving R9.4.1 the potential candidate of Nanopore sequencing-based HLA genotyping.

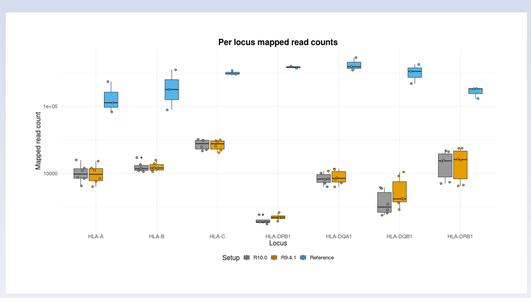

FIGURE 4 - Per locus mapped read count - The number of mappable reads is quite similar between the different Nanopore runs. The short-read reference data produced the highest number of reads, but further conclusions should be drawn carefully as the number of total reads are also different. Concerning Nanopore sequencing, HLA-DPB1 performed the weakest.

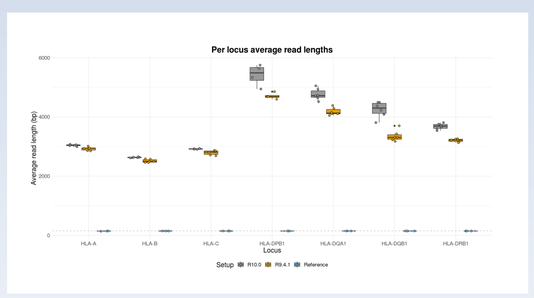

FIGURE 5 - Per locus average reads lengths - The plot depicts the difference between the short- and long-read sequencing data. Correlation between the amplicon and read length can be observed in case of the long-read data.

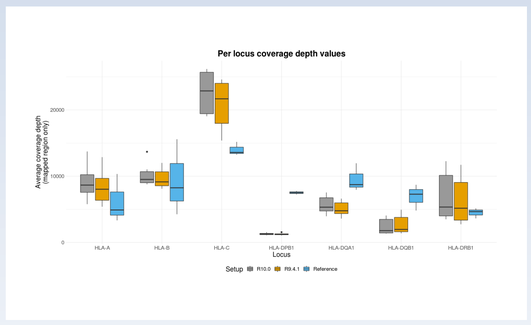

FIGURE 6 - Per locus average coverage depth values - Despite the significant differences observed in in the read counts, the average coverage depth values are more balanced, which can be explained by the read length differences. Besides the low read counts for HLA-DPB1, the average coverage depth values are also low.

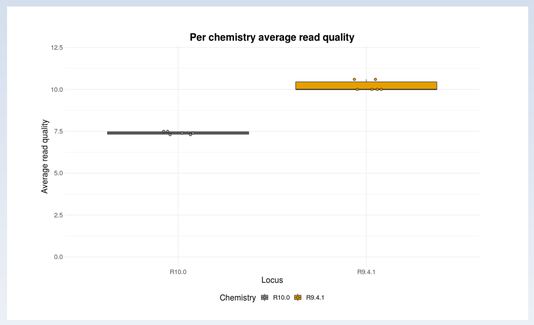

FIGURE 7 - Per chemistry average read quality - This plot depicts the differences observed in the average read quality values between the two different flow cell chemistries.

References

1. *Daniel, Branton. Nanopore sequencing: an introduction. World Scientific, 2019.*

2. *https://nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store*

3. *https://github.com/guigolab/bamstats*

4. *https://github.com/nanoporetech/pomoxis*

5. *https://github.com/nanoporetech/medaka*