The different pipelines produced consistent results. To compare the corresponding ones of the short-read ones, we analyzed the consensus sequences generated by HLA Twin with the same command. The 3-step pipeline handles both result generation and assessment, so the results can be directly compared between loci-specific primers. This way alignment artifacts present at the 5' and 3' consensus ends were excluded. The values were determined for read lengths, the coverage depth and read counts can also be assessed with the BAMStats tool. Because the full-length reference sequences were generated with majority rule. The result consensus sequences were then compared to the reference alleles of these 3 samples were prepared were prepared for Nanopore sequencing using the Ligation Sequencing Kit (SQK-LSK109) and the R9 chemistry, introduced in 2016. The different flow cell versions describe the nanopores used for long-read sequencing. The widespread introduction of Nanopore sequencing opened a whole new world for bacteria and eukaryotic genome sequencing with promises and challenges. 

Methods

A 24-sample, 7-focus library (HLA-A, HLA-B, HLA-C, HLA-DRBI, HLA-DQAI, HLA-DQBI and HLA-DRB1) was prepared using the Hyperion HLA v 2.74 /kit configuration and sequencing on Illumina MiSeq using the 300-cycle chemistry. These 3 samples are new, well-characterized cell lines (7MA, MOC, PGD) with known HLA genotypes were selected for Nanopore sequencing. The amplicons of these 3 samples were prepared for Nanopore sequencing using the Ligation Sequencing Kit (SQK-LSP109) and the PCR Barcoding Kit (SQK-PBAR01). 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 bow (v0.7.12-es190). Based on the alignments, consensus sequences were generated with majority rule. The result consensus sequences were then compared to the reference alleles and the consensus was calculated for each of them. The result alignment files were assessed with the BAMStats tool (v1.23) [5]. The short-read data was analyzed with Omixon HLA Twin 3.1.11 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 Ponseti (v0.3.13) [4] and Medaka (v1.11.1) [5] (Fig. 3). Finally, these consensus sequences were assessed with Ponseti and the result error profiles were compared. The plots summarizing the results were created with the ggplot library of R.

Results

The base-calling pipeline produces consensus correction 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 the 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 base calling output. While alignment artifacts are generated by bias, 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 were used for alignment generation, the consensus correction values were calculated considering only the target regions specified by the locus-specific primers. The new 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 Ponseti and Medaka can generate and analyze its 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 pipelines, this workflow iteratively corrects erroneous bases of Nanopore reads. The consensus assembly command of the Ponseti tool is able to determine consensus accuracy (Table 2) as well as both allele and sample-level filtering, indel substitution, insertion and deletion errors (Table 3). As 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.

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) than the average coverage depths differ to a lesser extent (Fig. 6). The read-lengths of Nanopore data also contribute with the amplitude length, 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 amplicon without any issue. The majority rule-based consensus sequences with the highest scores can be seen in cases 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 correction values for long-read data thus reach 99.50%.

When comparing the results of the combined pipeline of Ponseti 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 the 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 78,000 target reads 3-4 times slower. When one of R10.0 samples, the sequencing took approximately 17.5 hours to complete and reach target value.

Conclusion

The work presented here focus 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 sample for 7 HLA loci. Comparing the different types of data using alignments to known reference sequences reveals valuable information about the resolution of ions 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 reaching accuracy previously limited by the short-read approach. By comprising our data, assessing error profiles and differentiating random and systematic noise content we can develop better analysis methods and propagate the potential of Nanopore sequencing in everyday clinical routines.

Although the accuracy of consensus sequences of Nanopore data is still limited by the size of the short-read data, the analysis tools show promising results and they can be continuously updated 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 chemistry, a pore utilization problem is also observed during the sequencing. Conclusion between the employed error correction tools was created with the Pomoxis-specific sequencing platform is contrast to R10.0 leaving R9.4.1 the potential candidate of Nanopore sequencing-based HLA genotyping.