Detection of invalid DNA fragment splitting during RCA Nanopore sequencing

Charles Meijer *5938244* *30-01-2023* *Minor Research Project*

Abstract

Oxford Nanopore Technologies has recently established itself as one of the most popular suppliers of DNA sequencing machinery and protocols. However, the ONT pipeline is optimised to process relatively long strands of DNA, while its error rate is too large for it to be used to sequence shorter fragments. To work around this issue, CyclomicsSeq aims to create long DNA fragments containing many repeats of a short DNA insert. The DNA fragment is circularized through the addition of an optimised adapter and rolling circle amplification is used to produce the long DNA fragments containing alternating target DNA and adapter sequences for Nanopore machines to process. By creating consensus sequences from the many repeats of the same insert, CyclomicsSeq improves sequencing accuracy. However, the researchers noticed that the concatemerized fragment reads produced did not turn out as long as expected. Multiple factors could contribute to this phenomenon. The DNA fragments could break physically during sequencing, or the nanopore software could artificially split the reads. Fragment splitting could result in the loss of several repeats, leaving less data to build a consensus sequence, or cause one DNA fragment to be counted multiple times. This study shows that split reads can be identified by analysing information available in the sequencing summary provided by Nanopore and by aligning the reads to a reference genome. Features such as the presence of a nanopore adapter sequence, the delay between reads, alignment to a reference sequence, and the addition of barcode bases can be used to decide whether it is likely that two reads produced in succession were originally part of one DNA fragment. Wrongly split reads may be recombined so that the correct numbers of reads and repeats may be used in further analysis.

Layman's summary

DNA sequencing can provide insights into various diseases. It can reveal the presence of unwanted viruses or bacteria, or show that a cell could, or already has, become cancerous. Oxford Nanopore sequencing is a widely used DNA sequencing method. It is fast and cost-efficient, but it is not very well suited to sequence short DNA fragments. To combat this, CyclomicsSeq aims to create circular strands of DNA, which can be used to produce long DNA fragments, containing multiple repeats of the short DNA sequence, which can be used to form a 'consensus sequence' that has very high accuracy. However, while DNA fragments of 20k base pairs or more were generated, the average read length was only about half that size. This could be caused by the DNA being split, either physically, or by the sequencing software. This study shows that pairs of reads which were originally part of one larger fragment can be identified using information available in the sequencing data. Oxford Nanopore Technologies provides a summary of the sequencing experiment, which contains information on all the DNA fragments that have been read by the machine. Multiple features present in this summary file can be used to decide whether a DNA fragment was split or not. The presence of an adapter sequence, with which all fragments are supposed to start and

end, or the time that passes between two fragments being read can be used to identify fragments that may originate from one larger DNA fragment, that was split in two. When looking at the actual sequences, information like the specific barcode of each read, or overlap in the alignment of two reads to a reference genome can provide further evidence for DNA fragment splitting having taken place. This offers the possibility to recombine split reads so that the correct number of reads and repeats may be used during data analysis.

Introduction

Oxford Nanopore Technology (ONT) has recently become one of the most popular DNA and RNA sequencing techniques available, due to its speed, cost-efficiency, and flexibility in use (Deamer *et al.*, 2016). It does however have some limitations. The current sequencing pipeline is optimised to process relatively long DNA molecules, while it is less suitable for the analysis of shorter fragments.

CyclomicsSeq, a technique combining Oxford Nanopore sequencing with rolling circle amplification (RCA), was recently created to accurately sequence short DNA fragments (Marcozzi *et al.*, 2021). A DNA adapter, or 'backbone', is added to circularize the short 'insert' DNA sequence. Rolling circle amplification is used to generate long DNA molecules containing alternating insert and backbone sequences. Not only does this create long reads for which Oxford Nanopore sequencing is optimised, but by creating a consensus sequence of the insert and backbone repeats, read accuracy can also be improved (Marcozzi *et al.*, 2021). However, the analysis showed that the concatemerized DNA fragment sequences turned out shorter than expected. While 20kb+ fragments were produced by RCA, the average read length was only around 10kb. This could be caused by the DNA strands physically breaking into two or more pieces during sequencing, but also by ONT software splitting the reads artificially. The latter may happen if the DNA gets stuck inside the pore for example. The polarity of the pore can be reversed to make the fragment go in the opposite direction, and back out of the pore. Then, the polarity may be switched back so that a new DNA fragment can be used to create a new read. However, if the fragment that is stuck does not entirely leave the pore when the polarity is reversed, the sequencing of that fragment can be continued to create the next read. This way, multiple reads can arise from one DNA fragment. As CyclomicsSeq only uses reads with a minimum number of repeats for further analysis, read splitting may cause a usable fragment to split into two reads that would both be too short and are therefore thrown out. Also, one fragment may be counted multiple times after splitting if multiple reads, each containing enough repeats to pass the threshold, are produced. This would be especially undesirable if the technique is used to determine the relative amount of circulating tumour DNA, for example. BulkVis, a tool created to analyse bulk FAST5 files, is already capable of detecting reads that were incorrectly split in ONT data. However, it requires bulk data, containing not only the reads, but also the data generated when the pores are empty as input, and it relies on the alignment of the reads to a large reference sequence to operate (Payne, *et al.*, 2018). For the current CyclomicsSeq experiments, no bulk data was collected. Moreover, the alignment of reads cannot always be used, for example in the proof of concept of the CyclomicsSeq method, where only one

specific gene was targeted. For this project, the research question is: Is it possible to identify reads that are wrongly split after RCA sequencing? Identification of split reads could give more insight into the frequency of fragments being split, as well as an opportunity to recombine parts that the fragment split into, so the correct number of reads and repeats can be used during analysis.

Methods

For this project, in-house, readily available nanopore sequencing data was analysed using Python. The full notebook, containing all the code, is available via GitHub. A full list of the packages used for the analysis can be found there as well. Multiple features available in the summary and the alignment files were analysed. Start- and end times of the reads DNA were used to pair subsequent reads and calculate the time that passed between processing the two fragments. The start time of the template DNA of each read was used to infer the absence of the adapter sequence at the start of the reads. The alignment file contains the read-specific barcodes and the alignment of the reads to the reference sequence was used to find pairs of reads that mapped to the same region.

Results

Previous read

For the analysis, the focus lay mainly on the DNA fragments that followed each other through the same channel according to the summary file generated by the nanopore pipeline. The idea behind this is that if a fragment were to break, then its parts would be read one after another (Figure 1). If the fragment is split as a result of getting stuck in the pore and not being ejected completely during the polarity reversing process, the reads that came from the single original fragment would also be reported as if they were reads from multiple different fragments that followed each other closely through the same channel. Therefore, reads that were read in succession by the same channel were identified by sorting the all the reads produced by every single channel by their start times. All features that may indicate read splitting were tested for reads paired with the subsequent read that went through the same channel.

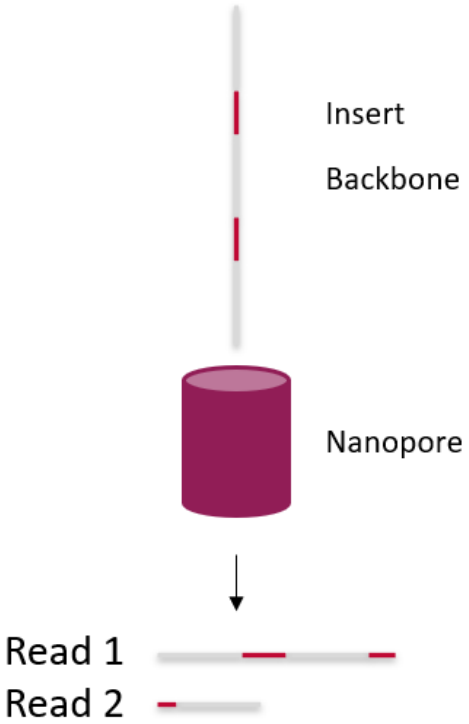

Figure 1. Schematic representation of read splitting during nanopore sequencing. A DNA fragment, containing repeats of backbone and insert sequences may be split during sequencing. When this happens, one fragment may result in two subsequent reads.

Nanopore adapter detection

To find pairs of reads that were originally part of the same DNA fragment, several features available in the summary and alignment files were analysed. Every DNA fragment processed by nanopore sequencing should theoretically contain nanopore adapter sequences at both the 5' and 3' ends. These sequences are added to the DNA to enable the fragment to interact with the pore's motor protein which moves the DNA through the pore to which it is attached. However, if the DNA fragment were to split, causing multiple reads to be created, the first part of the fragment should lack an adapter at the end, while the last part should lack an adapter at its start (Figure 2a). Therefore, finding a read that lacks an adapter sequence should point to that read being one of multiple parts of a single DNA fragment.

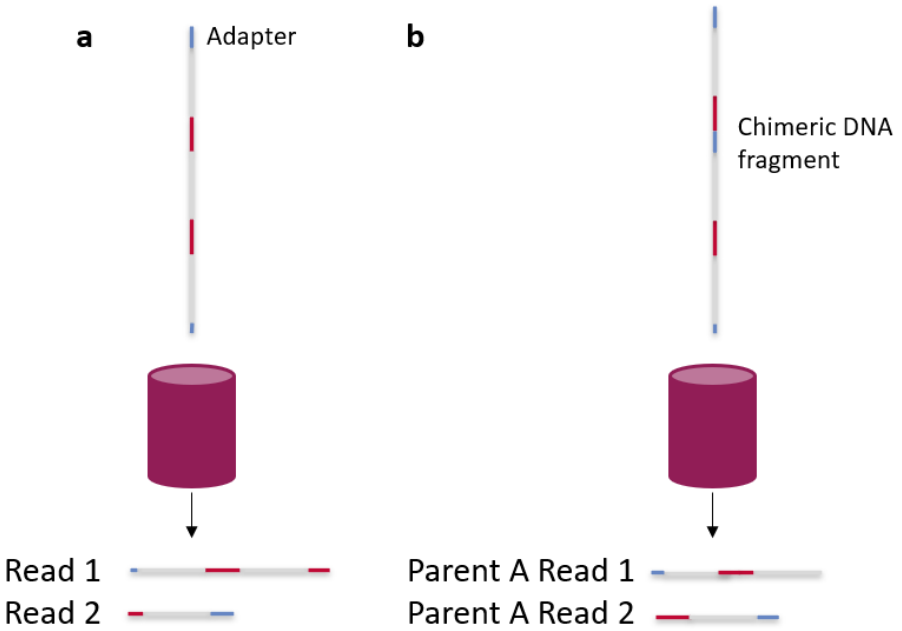

Figure 2. Adapter position can show read splitting or chimerization. Each fragment should contain known nanopore adapter sequences at both ends. (**a.**) Fragment splitting can cause reads to lack the adapter sequence on at least one end. (**b.**) Chimerization can cause an adapter to appear in the middle of a read. The nanopore software will split these chimeric reads into two.

The presence of the adapter sequence was inferred from the difference between the time when the DNA fragment would start going through the channel (start_time in the summary file) and the time when the actual template, containing the repeats of insert and backbone sequences, started being processed (start_time_template). If these two time values are the same, it would suggest that the template sequence would be read immediately upon the fragment entering the pore, which means the fragment did not contain an adapter sequence at the start. However, upon inspection of the raw data, which is the current change detected by the nanopore machine (Figure 3), it appeared this method of seeking out reads without an adapter sequence may not be reliable enough. It sometimes looked like the pattern usually showing the adapter sequence being processed was visible, even though the summary file would suggest it should not be there. The presence and location of the adapter sequence could also be tested by aligning the adapter sequence to the read and checking their similarity. ONT's base calling algorithm 'guppy' is configured to remove the nanopore adapter sequences by default. After base calling with this feature disabled,

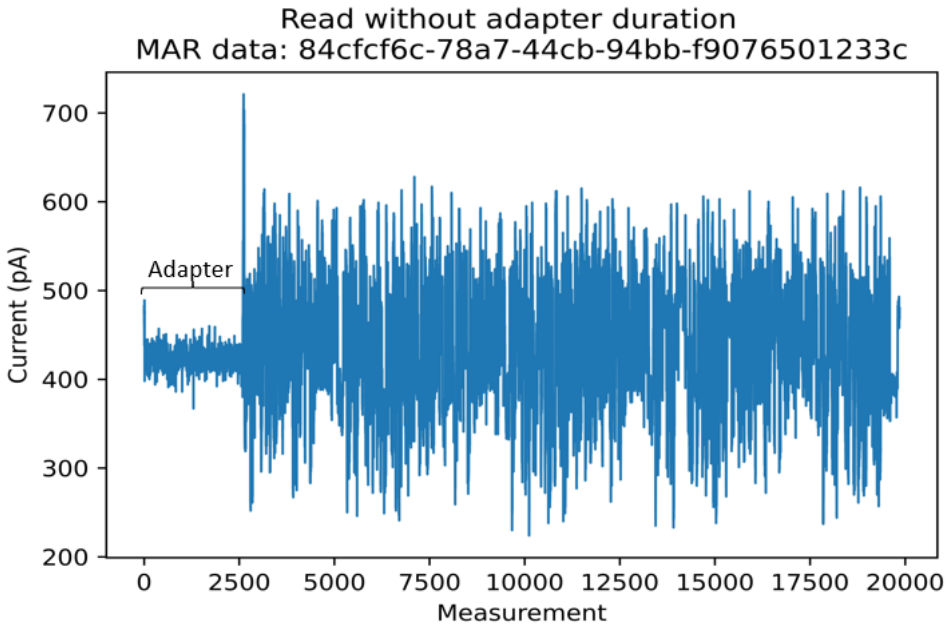

Figure 3. Raw nanopore data. Current change over time of a read that should not have an adapter, judging by the difference between start_time and start_time_template in the summary file. Multiple reads like these were found, which indicates that determining the presence of the adapter via this method may not be reliable.

analysis of the sequences revealed that reads would often start and end with only parts of the adapter sequence. This may be a result of it being difficult to accurately base call the beginning and end of the DNA fragments, which could cause the first and last couple of bases of the adapter sequences to not be present in the read.

Contrary to lacking an adapter, a DNA fragment could also contain more adapter sequences than expected. Chimerization, the fusion of different fragments, could result in one long fragment containing an adapter somewhere in the middle, as well as at both ends (Figure 2b). Nanopore's base calling software automatically seeks these reads out and splits them by removing the extra adapter sequence from the read. Each part is reported as an individual read. Leaving the reads of chimeric DNA fragments intact not only causes dissonance in the number of reads versus the original number of fragments but could also result in the creation of reads containing repeats of different inserts when performing whole genome sequencing. Creating a consensus out of these reads may lead to a nonsensical sequence, as well as the loss of the individual DNA fragments.

Time between reads

The second feature which may indicate fragment splitting is the delay between two sequential reads. It can be calculated by subtracting the 'end_time' of the first read from the 'start_time' of the next read from the same channel. If the time between reads is close to zero, it indicates that the reads followed each other very closely through the channel. This would also be expected from two reads that originated from one split DNA fragment, since the second part of the fragment would be immediately available to be sequenced when the first part has ended. These 'time between reads' values turned out very small (Figure 4).

While the time between reads is normally expected to be around 2-3 seconds on average, values ranging from 0 to 1 were most commonly observed for this dataset.

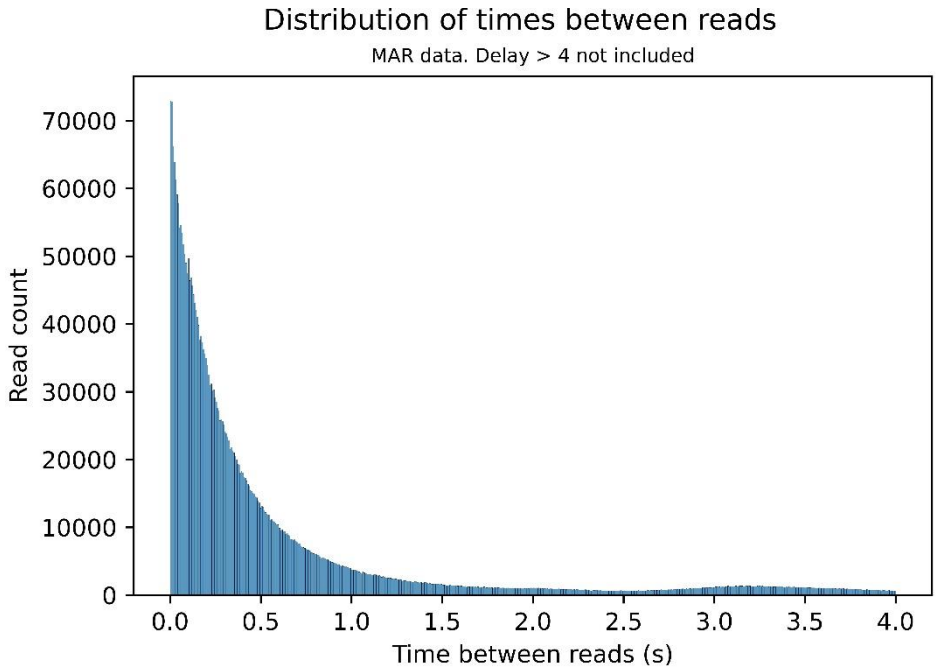

Figure 4. Distribution of delay between reads. Histogram depicting the distribution of the time that passes after one fragment is done being read and before the next one begins. The delay was very small, with most values being close to zero and only a small bump around 3.25 seconds.

The time that passed before each read was calculated for five different datasets: MAR, MARwg, WES, VER, and PAG. These datasets came from different experiments where nanopore sequencing was used to sequence DNA fragments. The datasets contained varying levels of repetitiveness in DNA sequences. MAR and MARwg both used CyclomicsSeq, which means that both samples contained a high percentage of backbone sequences. Furthermore, the MAR experiment was designed to determine a specific mutation in a known sequence, while MARwg used CyclomicsSeq to generate sequences from a whole genome. WES contained adapter sequences flanking each insert and VER contained a bacteriophage genome sequenced with ONT. More repetitiveness seems to result in a smaller delay between reads (Figure 5). This may be caused by the repetitive sequences enabling different DNA fragments to form structures and thereby keeping them together. This could result in the fragments being able to follow each other more closely through the pores, which would explain the difference in the time between reads among the different datasets.

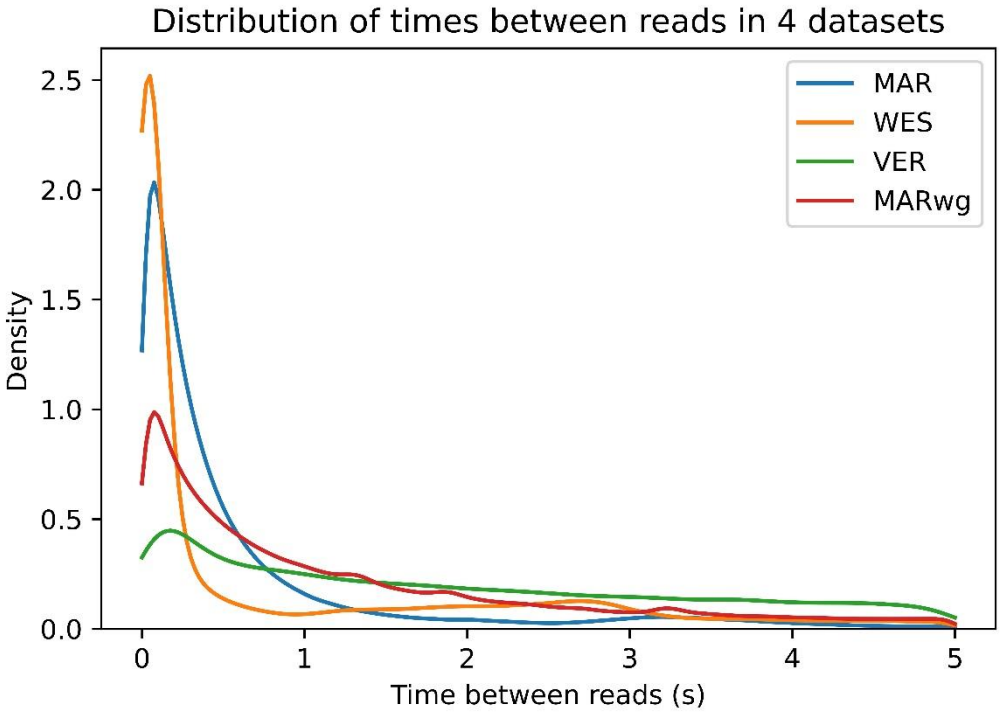

Figure 5. Distribution of times between reads for 4 different datasets. For MAR and WES, most of the times between reads were very close to zero seconds, with a small increase in density around 3-3,5 seconds. VER and MARwg, both containing fewer repetitive DNA fragments, show more evenly distributed values. The PAG dataset was left out, as its values were artificially skewed towards zero due to it containing many chimeric reads.

Since the delay between reads was generally small across ONT datasets, having just a little delay between reads does not necessarily mean that the reads were split from each other. Small times between reads could be a factor in identifying broken fragments, but should not be used as a definitive tool by itself. Having more repetitive sequences may also stimulate read chimerization. The PAG dataset, containing many repetitive oligonucleotides, showed a high number of chimeric reads, split into multiple reads by guppy when it detected the presence of the adapter sequence within the template sequence. The second part of the chimeric read automatically gets assigned a delay of 0.0 seconds, which caused the distribution of the times between reads of the dataset to be heavily skewed towards zero, and thus not very informative. However, as the oligonucleotides were marked individually, the high number of chimeric fragments did not form a real problem for the intended experiment. The delay being smaller during experiments with more repetitive DNA could even be an extra advantage of CyclomicsSeq. Less delay between reads naturally means that more DNA can be sequenced per unit of time, which would mean that CyclomicsSeq could provide an increased efficiency compared to regular nanopore sequencing if fragment splitting and chimerization can be dealt with.

Barcodes

In RCA nanopore experiments, the backbone sequences are given four spots for random bases. These barcode bases can be used to identify backbone sequences that were derived from the same plasmid, as these should all have the same sequence and therefore, the same barcode base combination. RCA uses more than 256 plasmids with random barcode sequences in the backbone, so more than the number of 4-base combinations. This means that the chances of two random reads containing the same barcode, even though they came from different plasmids, are slim, but not zero. Even so, the barcode sequences can still be a valuable contributor in making a case for two reads originating from one DNA fragment. However, it is difficult to retrieve each 4-base code after base calling. For the MARwg dataset, the full barcode was only found in ~21% of reads (Figure 6). Some barcodes could be partly recovered, but for ~55% of reads, no barcode base could be identified. About 2% of reads showed different barcode possibilities, which may indicate the presence of multiple different barcodes in those fragments. Another problem with using backbone barcodes is that even if they could all be retrieved, they are currently only analysed after consensus calling. This means that if a DNA fragment was split into two parts that are too small to be selected for consensus calling, their barcodes will remain unknown and cannot be used to identify and fix the split fragment. It would therefore make sense to try and retrieve the barcodes before consensus calling, but that might come at the cost of accuracy.

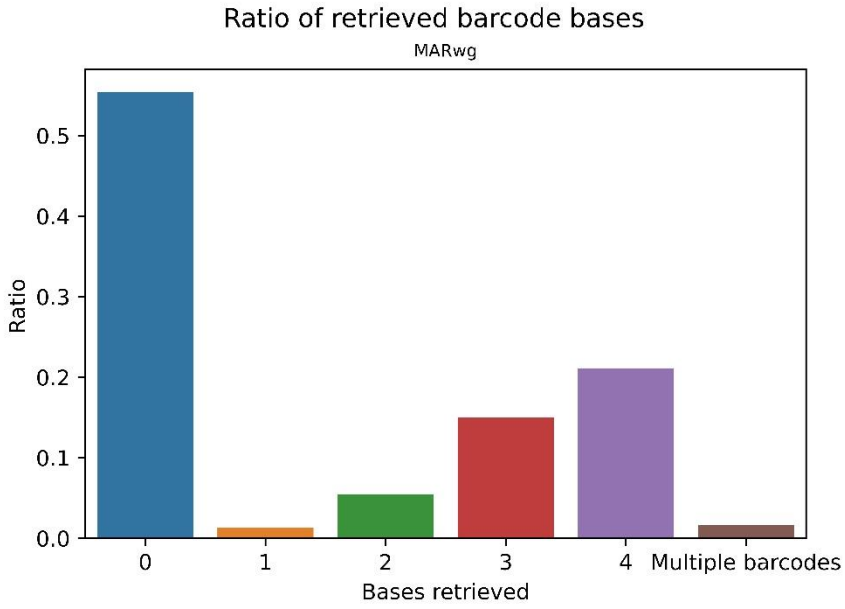

Figure 6. Number of barcode bases retrieved for MARwg. For over half of the total number of reads, not one barcode base could be identified. The full barcode was retrieved for about a quarter of the reads, while a small number of reads gave multiple possible barcodes, due to there being a difference between backbone repeats of that read.

Alignment to a reference sequence

The last feature analysed during this study is the alignment of the reads to the reference sequence (Figure 7). This information is not always available. For example, when targeting a single gene, all the reads can only align with either the backbone

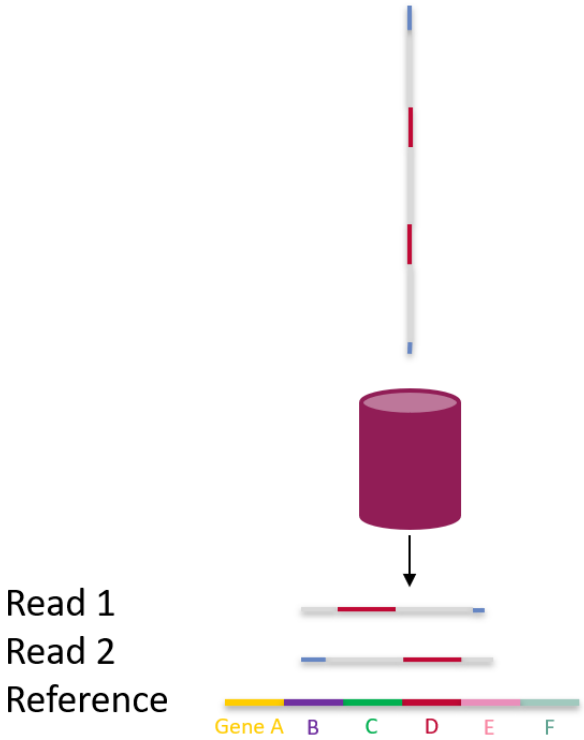

Figure 7. Schematic representation of 2 parts of a split DNA fragment mapping to the same reference region. In this example, there are two subsequent reads that both align to the same gene 'D'. The chances of this happening become less and less likely with a larger reference genome.

sequence or the gene of interest. However, when performing whole genome sequencing using RCA, the chances of two random reads aligning to the same region of the genome become much slimmer. Therefore, finding a pair of subsequently produced reads that map to the same region in the reference sequence, may indicate that those reads originally belonged together. It is also important to check whether the two reads have the same orientation. It may be possible that the forward strand just closely followed its reverse counterpart through a channel, which is something that ONT stimulates with its duplex read technology (Oxford Nanopore Tech Update, 2021). Similar to CyclomicsSeq, the idea is that the forward and reverse reads can be used to create a consensus sequence and thereby gain a higher sequencing accuracy. A sample of reads of the MARwg set was analysed to identify reads that were produced in succession by the same channel, aligned to the same region in the reference genome, and had the same orientation. These criteria are very specific, and the pairs of subsequently created reads rarely matched them. Only ~1% of reads aligned to the same region as the read that was processed before it, and only ~0,4% also had the same orientation. These numbers do not represent every incorrectly split DNA fragment, as fragment splitting could also sometimes result in reads that are too short to accurately map to the reference genome, which means that it cannot be identified via alignment. Even so,

this analysis suggests that fragment splitting may only occur sporadically, and will probably not form a true problem for the experiment, given a large enough dataset.

Discussion and Conclusion

Fragment splitting can cause one DNA sequence to be counted multiple times or create reads that do not meet the minimum number of repeats, which would result in them not contributing to the analysis at all. Several features found in the summary and alignment files can indicate pairs of reads that resulted from DNA fragment splitting. In theory, the presence and location of the adapter sequence could aid in identifying split reads, since it should be present at the start and end of each read. In practice however, it proved difficult to reliably find reads that lacked an adapter, as the information in the summary file does not always line up with the pattern of the voltage change that is detected when reading the DNA fragment. It also appears that the start and end of the fragment may not always be accurately sequenced, as disabling ONT's base calling software from automatically removing the adapter sequences from the reads would often result in reads starting and ending with only parts of the adapter sequence. The same problem was encountered with the barcode bases. Like the adapter sequences, the barcodes can theoretically be useful in identifying fragment splitting, as the odds of randomly finding a pair of reads in sequence with the same barcode are slim. But like the adapter sequences, the barcodes cannot reliably be used at this moment. None of the barcode bases could be identified in over half of the reads in the MARwg sample with our current tools, and the complete barcode was only retrieved in less than a quarter of all reads. Moreover, the barcodes are currently identified in the consensus sequences for each read. This means that if two portions of a split DNA fragment are excluded due to insufficient repeats, no consensus calling will take place and their barcodes cannot be checked for similarity. On the other hand, if the reference sequence is long enough, aligning the reads and checking for read pairs that show similarity with the same region can be a powerful way to identify fragment splitting. After selecting read pairs that map to the same region, the orientation of both reads can be checked to see if the reads were truly the same, or if the forward sequence was quickly followed by its reverse counterpart. The delay between reads could be used as an extra factor in providing evidence for fragment splitting, but as the delay values were all generally low in CyclomicsSeq experiments, it does not hold much value on its own. It is currently difficult to determine which reads originally belonged together when spatial information is not available. If the adapter sequence can be identified reliably and the barcode bases can be retrieved more frequently, then together with the orientation of the reads, there may be enough evidence to show that fragment splitting did occur within an experiment with a targeted approach. And even if broken DNA fragments could be identified, it leaves the question of whether it is worth the extra time and resources. In the MARwg whole genome sample, only ~1% of reads aligned to the same region as the previous read processed by the same channel, and only ~0,4% also had the same orientation. This means that the problem appears to be less significant than previously anticipated. So, while some reads that originated from one

fragment can be identified and recombined, doing so should generally not offer significant contribution towards the goal of the experiment.

References

- Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature Biotechnology*, *34*(5), 518–524. https://doi.org/10.1038/nbt.3423
- Marcozzi, A., Jager, M., Elferink, M., Straver, R., van Ginkel, J. H., Peltenburg, B., Chen, L. T., Renkens, I., van Kuik, J., Terhaard, C., de Bree, R., Devriese, L. A., Willems, S. M., Kloosterman, W. P. & de Ridder, J. (2021). Accurate detection of circulating tumor DNA using nanopore consensus sequencing. *npj Genomic Medicine*, *6*(1). https://doi.org/10.1038/s41525-021-00272-y
- Oxford Nanopore Tech Update: new Duplex method for Q30 nanopore single molecule reads, PromethION 2, and more. (2021, 3 december). *Oxford Nanopore Technologies*. https://nanoporetech.com/about-us/news/oxford-nanopore-tech-update-new-duplex-method-q30-nanopore-single-molecule-reads-0
- Payne, A., Holmes, N., Rakyan, V. & Loose, M. (2018). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics*, *35*(13), 2193–2198. https://doi.org/10.1093/bioinformatics/bty841