Introduction

Cpf1 is an enzyme which can be used in CRISPR systems to give a cut in the target sequence when a guide sequence is given. Two families of Cpf1, AsCpf1 from (*Acidaminococcus bacterium*) and LbCpf1 from (*Lachnospiraceae bacterium*) have been shown [1] to result in successful genome editing in human cells with FnCpf1 also been used more often in genome editing in plants.

Cpf1 recognises a thymine rich PAM distal to the cut site with cleavage resulting in a staggered cut. The exact position of the cleavage sites on both strands of the DNA and therefore also the length of the overhang is not yet fully understood; most commonly the cleavage has been observed 18 nucleotides after the PAM on the non-target strand of DNA and at the 23rd nucleotide in the target strand, resulting in a 5nt overhang [2]. However, it has also been seen to induce 2-4nt overhangs around the same position [3]and more recently that the length of the spacer sequence may influence the cleavage position on the non-target strand, with shorter (17-19nt) guide sequences giving cleavage after the 14th nucleotide on this strand, in addition to the 18th nucleotide, resulting in longer overhangs [4]

This work uses some of the data given in [5] to look into several properties of the indels produced including: distributions of the lengths and positions of the indels, the occurrence of microhomologies and their effect on the deleted regions and insertions.

Results

## Range of activity of Cpf1:

*Across the 1251 guide-matched target sequence pairs a range of Cpf1 activity from 0 to 100% was observed and there was reasonable agreement with the results found in* [5]*. Some indels occurred frequently across a range of guides.*

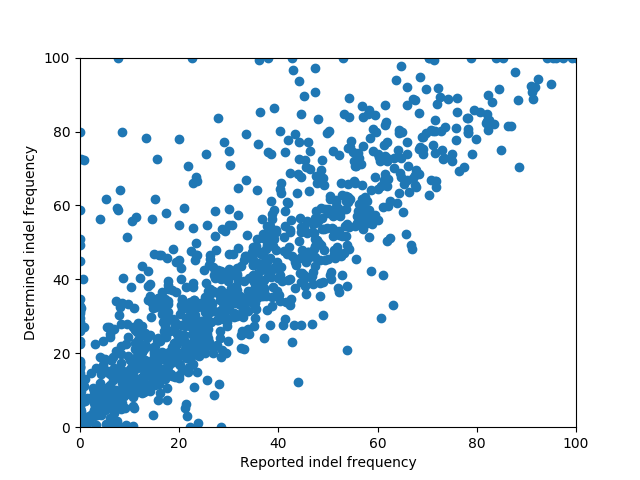
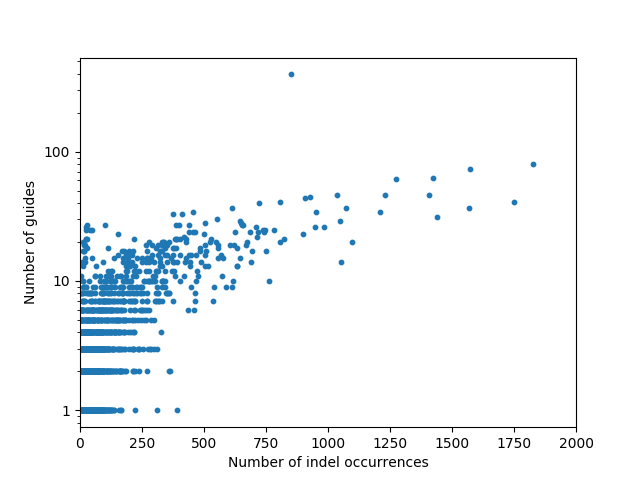


Figure 1 The percentage of indels formed after 5 days for each of these 1251 guides and matched target pairs used in the study [5] was computed and compared against the indel frequencies previously reported. The two percentages were correlated (Pearson correlation coefficient of 0.87) with the indel frequencies determined using this method often higher than those in the previous study. Possible reasons for this include: introducing a bias as a result of the filtering process where reads for some guides were not included in the analysis, not subtracting the background indel frequency in the cell library, as done previously or, more generally, differences in the methods used to detect indels or classify single-base substitutions. A range of activity levels from 0 to 100% across the guides was observed suggesting that, like Cas9, there are factors involved in the efficiency of Cpf1 cleaving target sequences. Activities from indels were called assuming that the cut position was at 18 and the ‘max\_cut\_dist’ was 5.

# **Indel lengths**



|  |  |  |
| --- | --- | --- |
| Indel | Number of Guides | Number of indel occurrences |
| D6 L-6R1 | 91 | 1830 |
| D1 L-8R-6 | 34 | 1666 |
| D1 L-4R-2 | 34 | 1638 |
| D1 L-6R-4 | 40 | 1587 |
| D5 L-6R0 | 76 | 1573 |
| D7 L-7R1 | 61 | 1426 |
| D2 L-9I2C13R5 | 376 | 632 |

Table 1: This table shows indels of particular interest seen in figure 2. They may either appear across a large number of guides or have a high count of number of occurrences. The first 6 deletions relate to the darker coloured squares in Figure 8.

Figure 2: With the number of indel occurrences normalised so that each guide had the mean read count of 491, this plot gives the total number of times individual indels occurred across the guides that each indel appeared in. This was used to confirm that the results were not biased to a great extent due to indels being observed frequently across a small number of guides and to check how evenly distributed the indels were.

Cpf1 cutting results in indels with a range of indel lengths. Long deletions are common, with a preference for those with length 1 and 6. Insertions occur far less frequently, which contrasts with the behaviour observed with Cas9.

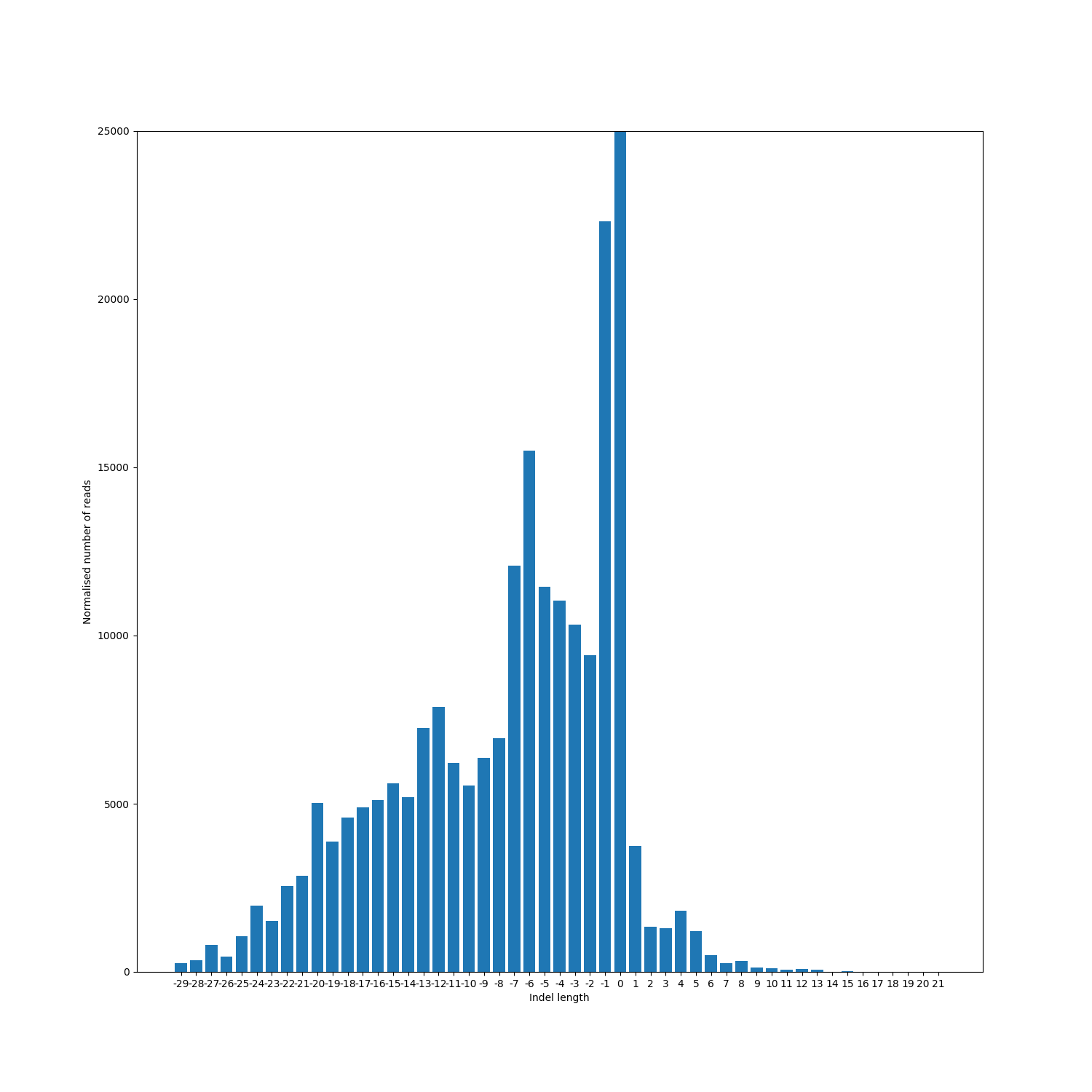


Figure 3 Profile of the net indel length observed across the 1131 guides which have a read count above the threshold of 100. Normalising each guide to have the mean read count, 491, reduces bias from guides with high and low read counts. For reference, the height of the bar corresponding to no net change in length is 364 408 which corresponds to around 65% of indels, and this includes times when no cut was made or indels where base substitutions were made. From this one observes that deletions occur far more frequently than insertions with single base pair deletions and deletions with length 6 observed more often. The notable drop in frequency between deletion lengths 7 and 8 may be indicative of a change in type of the deletions, for example by going from a deletion of both overhanging regions for shorter ones to deletions between microhomologies for longer deletions.

### Change in profiles across different guide activity levels

|  |  |
| --- | --- |
| 1 | 0 – 13.1% |
| 2 | 13.1 – 31.1% |
| 3 | 31.2 – 52.9% |
| 4 | 52.9 – 100% |

Table 2: Separating the guides according to indel frequency gives the following 4 groups

# 

Figure : Percentage of indels of each length associated with each set of guides with different activity level to determine if some indel lengths were favoured by the more or less active guides with respect to the others. We see that the shorter insertions are less favoured by the most active guides compared to how frequently deletions across a range of lengths were observed , whereas for the less active guides these insertions occur in roughly the same proportion as deletions.

# Deletions

### Microhomologies

*The longer the microhomology the more likely there is to be a deletion between them (?). When the effect of repeats on the proportion of deletions of each length is considered, repeats of length 3 have the most influence for longer deletions and since the proportion of deletions associated with no repeat or a repeat of length 1 decreases with deletion length, longer deletions are more target-specific.*

Table 3: Count of the number of the repeats (microhomologies) of lengths 2 to 9 occurring in the target sequence across all 1251 guides and the number of deletions that were between and including one set of the repeated bases. This was expressed as a percentage of the total number in the third column. When the microhomologies are filtered so as to occur on either side of one of the proposed cut sites (14 or 18) the bracketed numbers give the respective counts. The filtering was performed to give a more indicative total number of repeats that might have been expected to give a deletion to remove potential cases of adjacent repeats away from the cut site. Although across all of the guides microhomologies with length greater than 5 were not that frequent, the percentage of number of deletions due to microhomologies of this length compared to the number of such microhomologies increases with length, suggesting that the longer the microhomology the more likely there is to be a deletion between them, although there is insufficient data to justify this.

|  |  |  |  |
| --- | --- | --- | --- |
| Microhomology/ repeat length | Total number of microhomologies or repeats | Number of deletions from microhomologies or repeats | Percentage |
| 2 | 70 801 (45 797) | 2089 (1693) | 3.0 (3.7) |
| 3 | 14 820 (10 511) | 932 (807) | 6.3 (7.7) |
| 4 | 3 640 (2 635) | 320 (283) | 8.8 (10.7) |
| 5 | 947 (686) | 116 (107) | 12.2 (15.6) |
| 6 | 225 (164) | 35 (35) | 15.6 (21.3) |
| 7 | 53 (39) | 9 (9) | 17.0 (23.1) |
| 8 | 14 (10) | 3 (3) | 21.4 (30.0) |
| 9 | 4 (2) | 0 (0) | 0.0 (0.0) |

Oligo 846: TTTACTCACAGTAACAAGACTACTCCCAGCTTGGCGTAACTAGATCT

D25 L-13C5R18: TTTACTCACAGTAAC TAGATCT

D11 L-7C2R7: TTTACTCACAGTAACAAG CTTGGCGTAACTAGATCT

D18 L-15C3R7: TTTACTCACAG CTTGGCGTAACTAGATCT

D12 L-12C1R2: TTTACTCACAGT CCaAGCTTGGCGTAACTAGATCT

D7 L-6R2: TTTACTCACAGTAACAA CCCAGCTTGGCGTAACTAGATCT

Figure 5: Each deletion may be classified, firstly, according to whether there is a microhomology or repeat associated with it and further by the length of this microhomology when it exists. This figure shows examples of deletions for the guide sequence 846 showing the sequence of bases that could be at either end of the deletion in red and the deletions would be classified into the groups 5,2,3,1 and 0 respectively according to the length of this repeat. The deletion D12 L-12C1R2 also presents a single base substitution, marked in green.

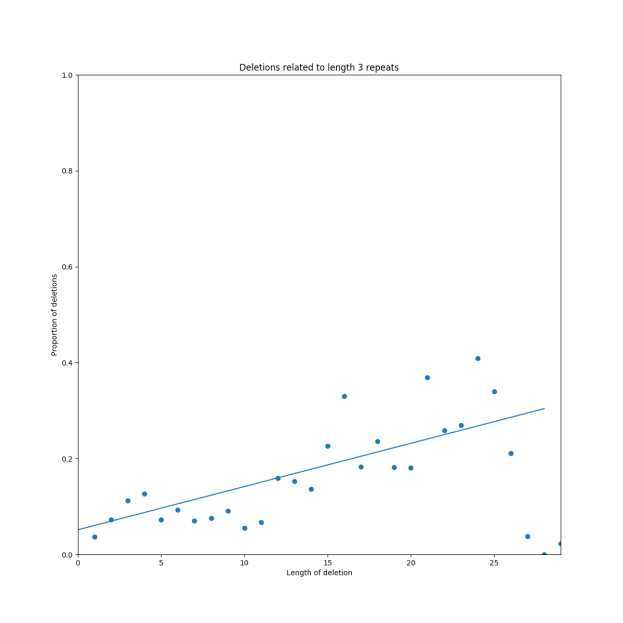
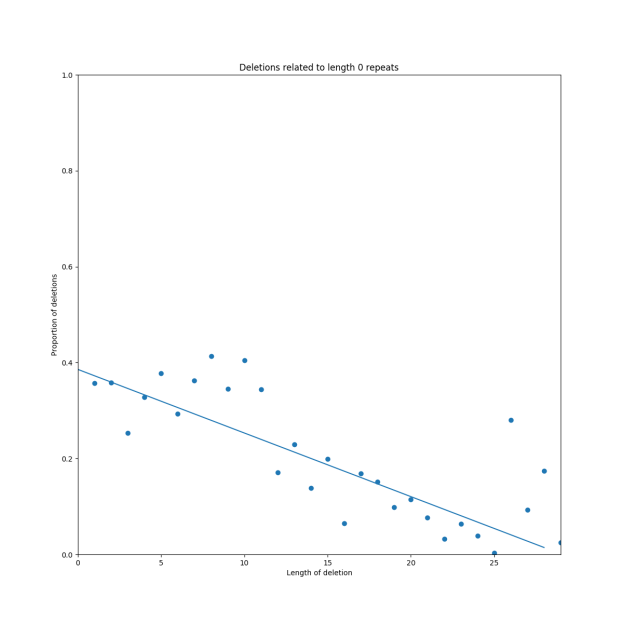
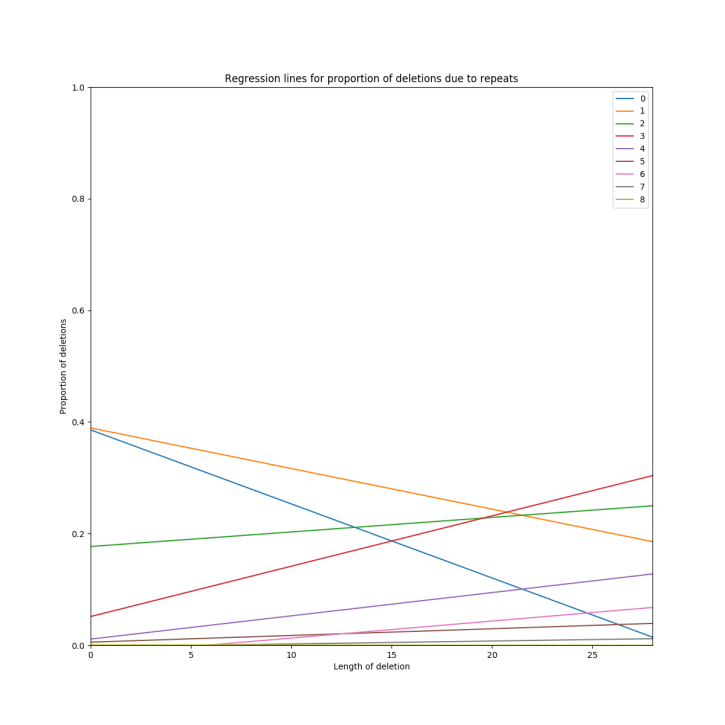


Figure 6: For each deletion length all of the corresponding deletions were found and classified in the way described in figure 4. Then, for each of these classified groups, the proportion of deletions of each length in this group compared to the total weight of deletions of that length was plotted as seen in the two plots on the left. Regression lines were added for each showing the relationship between the length of the deletion and the proportion of deletions due to repeats of that length and the figure above combines these lines in one plot.

As there are slopes with negative gradients when the deletion is between no repeating sequence or a repeat of length 1, whilst those lines for repeats of length at least 2 have positive slope, it suggests that longer deletions are more target specific with repeats of length 3 seemingly most influential. However, microhomologies with length greater than 5 were infrequently seen in the target sequences, with potentially different guide-target pairs, the effect on the proportion of deletions by these can be observed.

### Left and right positions of deletions

There are distinctive profiles for the positions of both the upstream and downstream ends of the deleted regions. The strong peak in the left (upstream) profile suggests Cpf1 cleavage predominantly at the 14th nucleotide of the target sequence whereas there may be more variability in the position of cleavage on the target strand. The two positions are not independent with particular indels preferred, indicating that deletions may frequently occur between the cleavage positions on both strands.

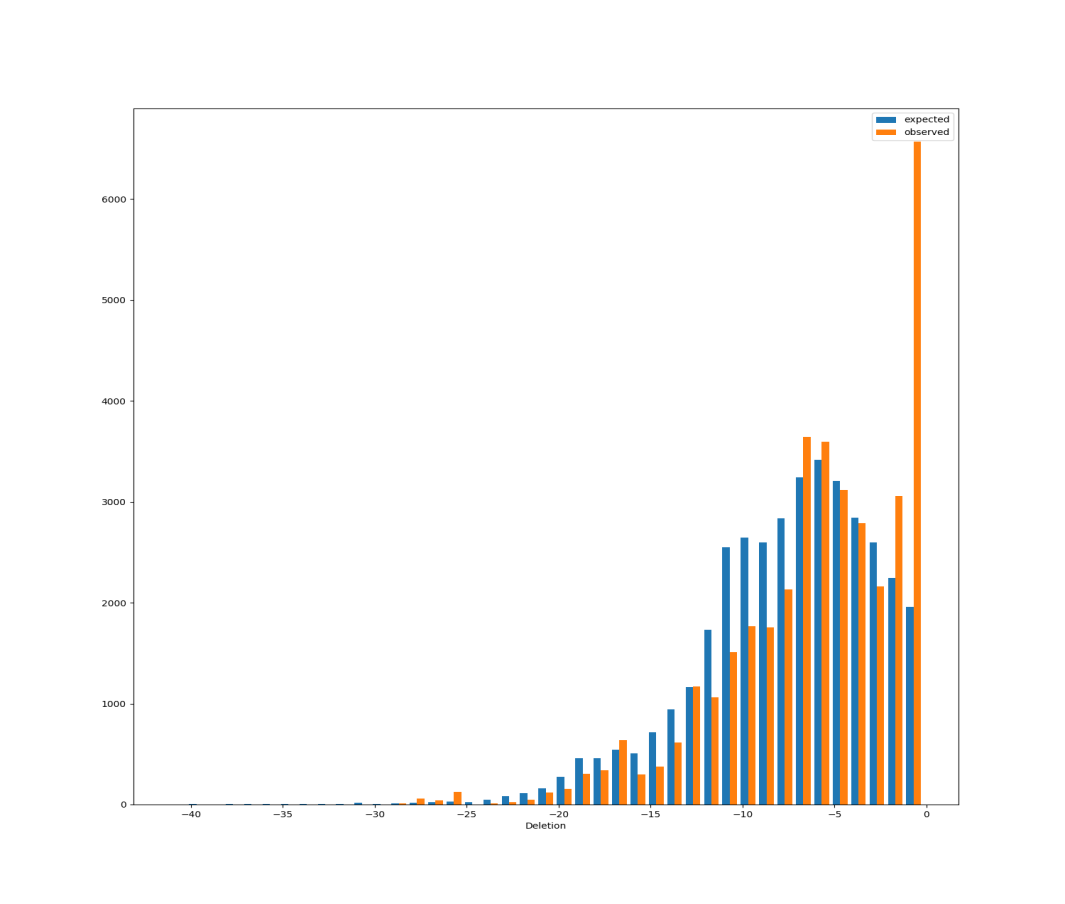


Figure 7: Given the two distributions for the profile of upstream and downstream positions of the deletions in figure 6, we can enquire as to whether they might be independent. Randomly choosing one upstream and one downstream deletion position independently with probabilities defined from these distributions and computing the corresponding deletion length, the profile of deletion lengths in blue is obtained and compared with the observed profile in orange.

The main differences between the two distributions are that deletions of 1 base pair occur far more frequently than the number of occurrences that might be expected if both positions were independent and deletions of length 8 to 11 occur less noticeably less frequently than may be expected.

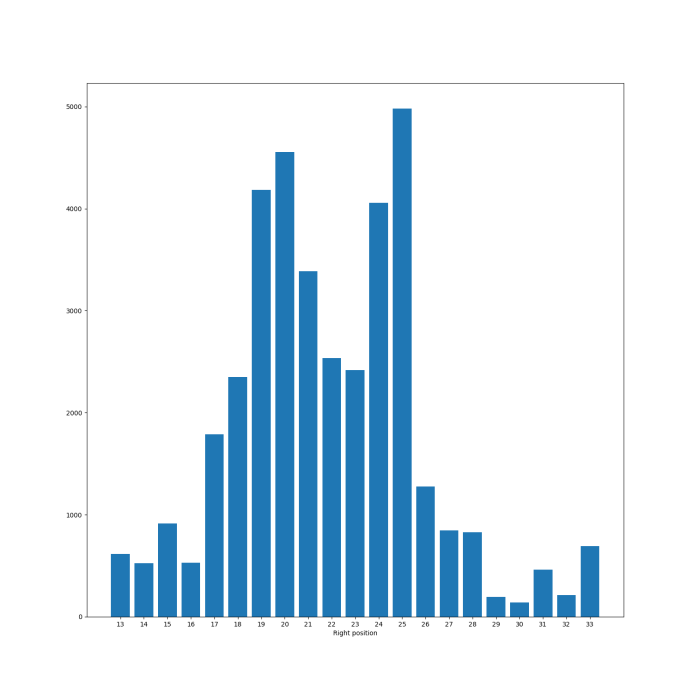
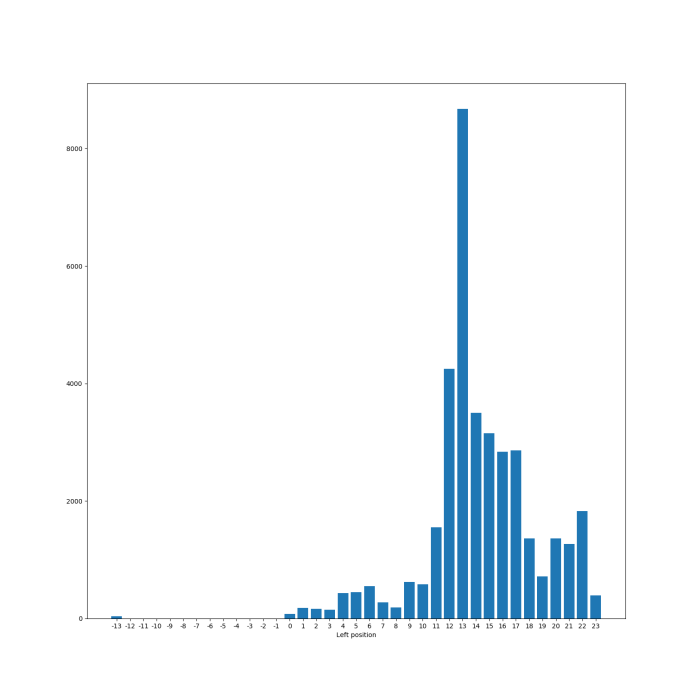


Figure 8: Profiles for the left and right positions of deletions that are not associated to microhomologies or repeats in the target sequence. The labels on the x-axis represent the number of nucleotides after the PAM that the last (left) or first (right) nucleotide that match in both the target and sequenced read after cutting. The strong preference for the left position to be at 13 seems to provide evidence that a large proportion of the time the cleavage site on the non-target strand was at the 14th position of the guide, perhaps as opposed to the 18th or 19th where the count of the number of related indels was much lower. In contrast, the profile for the position of the right side of the deletion is very different and might suggest that there is greater variation in the cleavage position on the target strand. The peaks at 19/20 and 24/25 provide evidence for two such positions, particularly as the frequency of deletions having right positions outside of the region between is much lower.

## 

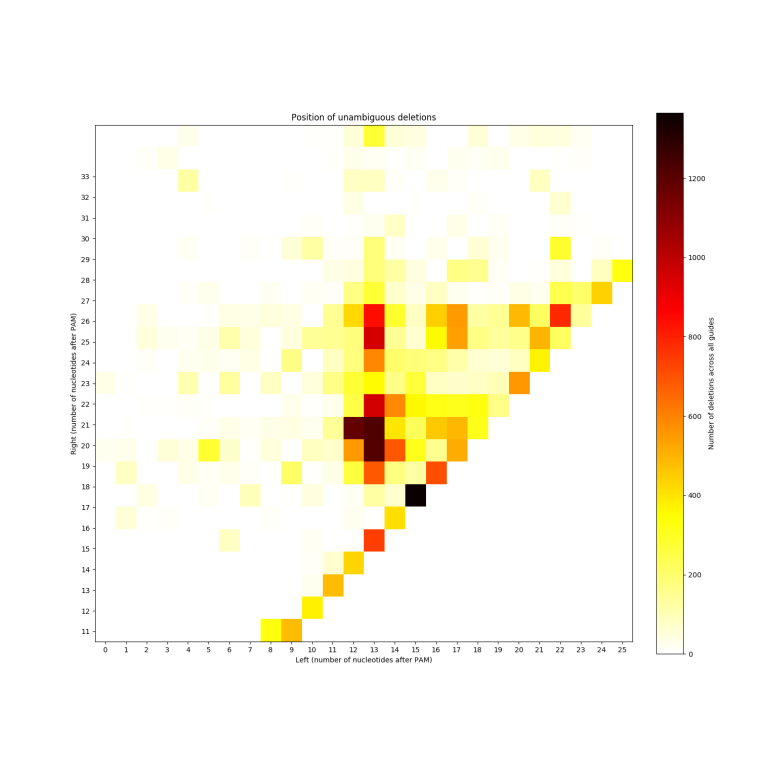


Figure 9: Heat map showing the counts for each pair of left and right positions across the deletions where there are no repeats or ambiguities associated with the lowest diagonal line corresponding to a deletion of a single base pair. Independence of both ends of the deletion would mean that a more uniform colour across the various rows and columns would be seen. However, this is not the case with some deletions, for example between the 14th and 20th nucleotides in the target sequence and of the 16th nucleotide, being greatly favoured over others. Again, this provides strong evidence that these deletions correspond to deletions around cleavage positions on both strands.

## Insertions

*Insertions occurred far less frequently than deletions with insertions of 1, 3 or 4 base pairs observed more commonly than other insertion lengths. The insertions tended to repeat bases towards the end of the target sequence possibly providing evidence for Cpf1 cleaving the target sequence around the 18th and 19th nucleotide on the non-target strand and giving on overhang of 3 or 4 nucleotides. A possible explanation for observation of this type of insertion is provided in Figure 12.*

1. TTTAACAGGAGGAGGTCGTCGTTGTT CAGCTTGGCGTAACTAGATCT

TTTAACAGGAGGAGGTCGTCGTTGTTTTGTTCAGCTTGGCGTAACTAGATCT

1. TTTAACAGGAGGAGGTCGTCGTT GTTCAGCTTGGCGTAACTAGATCT

TTTAACAGGAGGAGGTCGTCGTTCACGTTCAGCTTGGCGTAACTAGATCT

1. TTTAGCCCTCATCCTTGGTGGTGA AGTAGCTTGGCGTAACTAGATCT

TTTAGCCCTCATCCTTGGTGGTGATGAAGTAGCTTGGCGTAACTAGATCT

Figure 10: Examples of insertions. It was noted that the insertions of the form a. and c., with one set of the possible inserted bases being identical to a sequence of bases in the target sequence starting around the 18th or 19th nucleotide downstream of the PAM, were common. These insertions will contribute to the blue bar in Figure 9 and points at 18 for insertion sizes 5 and 3 in figure 11. respectively. Other types of insertion also occurred as in b., the inserted bases are not of this form so this insertion would contribute to the orange bar in Figure 9.

Figure 11: Insertions made up around 1.5% (18.44/1251) of the indels across all of the guides, with an insertion of a single base pair favoured (~34% of insertions) followed by insertions of length 4 and 3 (18% and 16% respectively). The frequency of insertions with length greater than 5 was much lower. The vertical scale weights each insertion according to the percentage it occurred out of all indels in a particular guide with each guide having a total weight of 1. The blue bar presents the times when the insertion has the form of the inserted bases repeating an adjacent sequence of bases already in the target sequence, with orange the remainder of insertions. The large proportion of blue seen in each of the bars gives plausible evidence for the outline of the mechanism shown in figure 12.

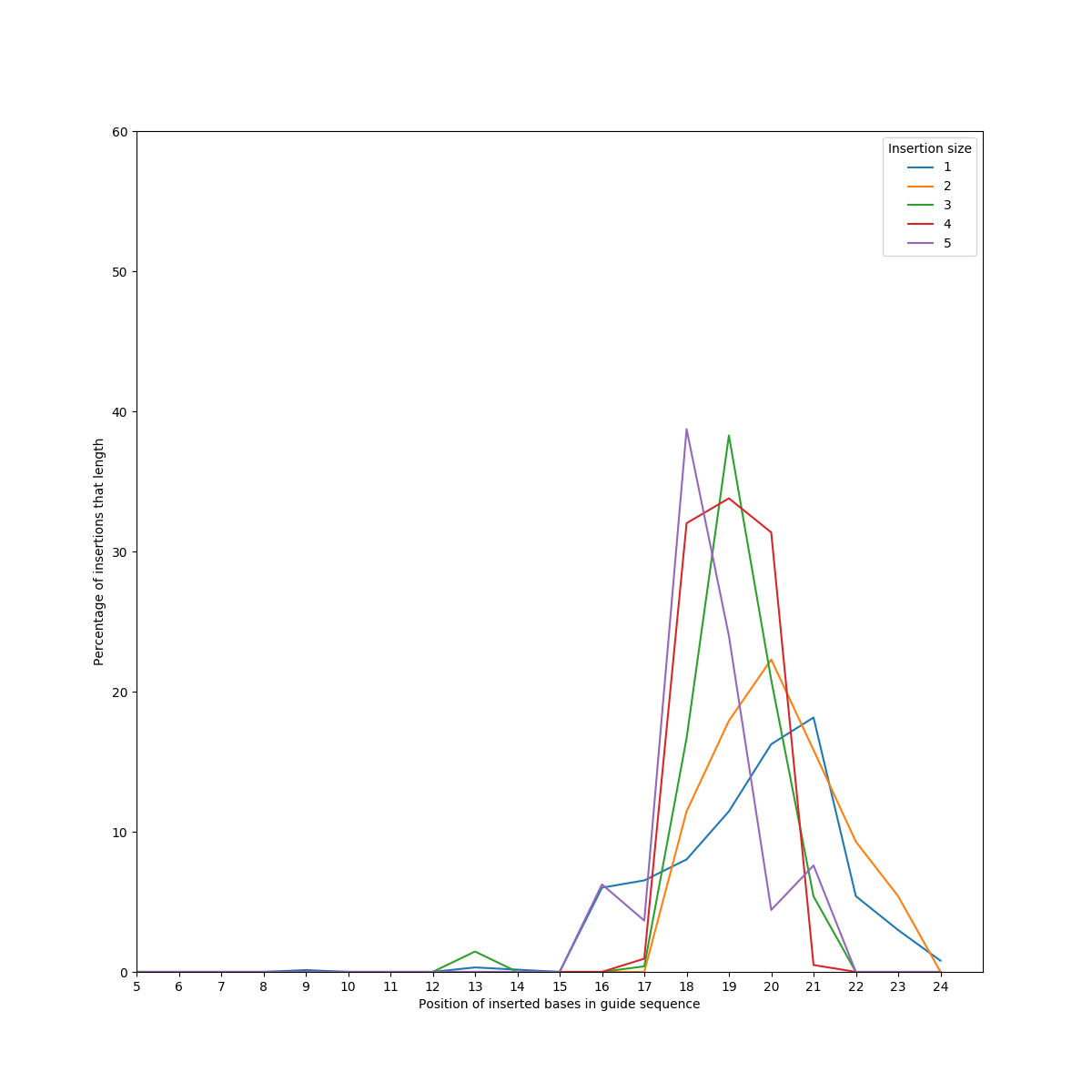
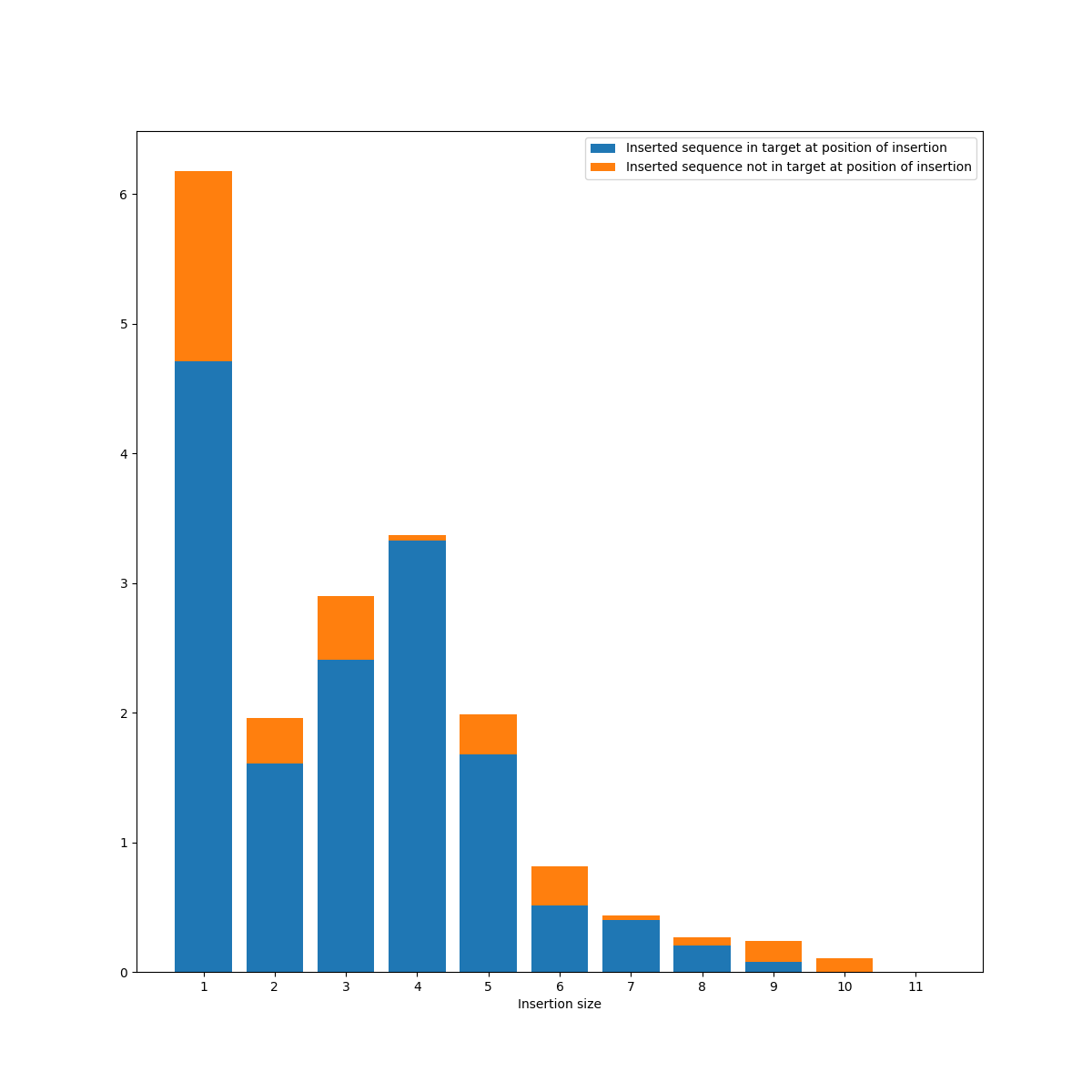


Figure 12 Plot showing the positions of the start of the bases in the target sequence that were repeated for insertion sizes 1 to 5, when the insertion had the form of a. or c. in Figure 10. When multiple positions could have been selected, one was randomly chosen from these. The greater percentage seen at the 19th position for insertion sizes 3 and 4 suggests that cleavage can occur at this position. Although there is more variation in the positions when the insertion size was 1 or 2, the modal position for both would correspond to the cleavage position on the target strand at the 22nd nucleotide. Far less frequently there were insertions of length 8 and 9 seen around positions 14-16 in the guide sequence which correspond to the cleavage site around the 14th base.

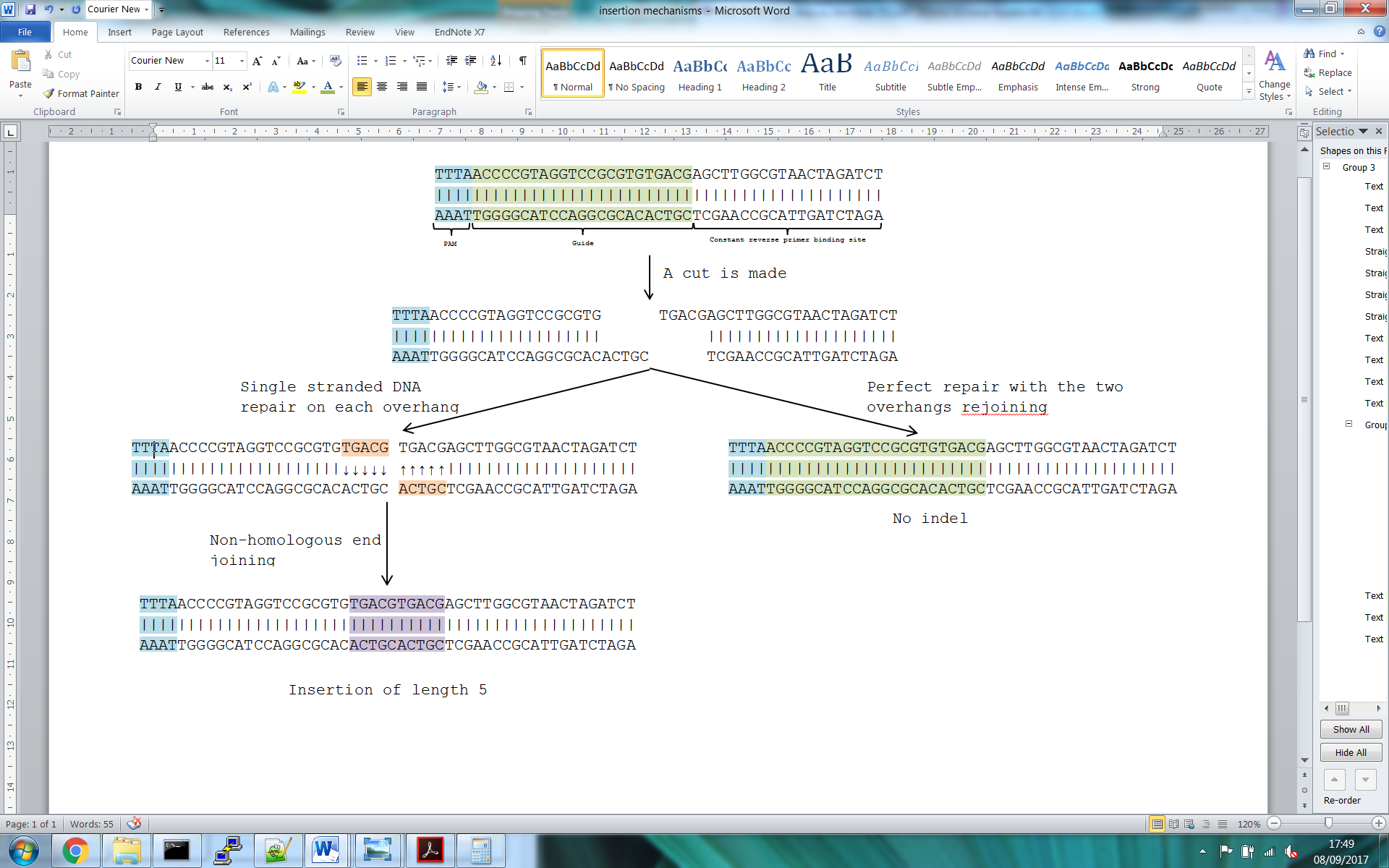


Figure 13: A proposed mechanism for insertions of this type with single strand repair for each of the overhangs followed by non-homologous end-joining. From figures Figure 9 and Figure 11, this might explain the shorter insertions.

## 

Figure : When the insertion repeats part of the guide sequence, this plot shows the positions of both ends of the inserted region, assuming that it occurs before the sequence in the guide for example, TTTAACAGGAGGAGGTCGTCGTTGTTTTGTTCAGCTTGGCGTAACTAGATCT is included in the box with the start of insertion at 18 and end of insertion at 22.

## 

Figure : When the insertion repeats part of the guide sequence, this plot shows the positions of both ends of the inserted region, assuming that it occurs after the sequence in the guide for example, TTTAACAGGAGGAGGTCGTCGTTGTTTTGTTCAGCTTGGCGTAACTAGATCT is included in the box with the start of insertion at 23 and end of insertion at 27. This plot shows that the start of the insertion most frequently occurs at the 23rd nucleotide suggesting that the cleavage position on the target strand is at the 22nd nucleotide.

## Cleavage on non-target strand in multiple positions

Oligo 1023: TTTAGCCCTCATCCTTGGTGGTGAAGTAGCTTGGCGTAACTAGATCT

D5 L-6C1R1: TTTAGCCCTCATCCTTG GAAGTAGCTTGGCGTAACTAGATCT 8.73%

D7 L-7R1: TTTAGCCCTCATCCTT AAGTAGCTTGGCGTAACTAGATCT 5.66%

D12 L-7C1R7: TTTAGCCCTCATCCTT GCTTGGCGTAACTAGATCT 6.60%

D2 L3R6: TTTAGCCCTCATCCTTGGTGGTGAAG GCTTGGCGTAACTAGATCT 4.72%

D4 L0C1R6: TTTAGCCCTCATCCTTGGTGGTG AGCTTGGCGTAACTAGATCT 3.78%

I3 L-2C3R2: TTTAGCCCTCATCCTTGGTGGTGATGAAGTAGCTTGGCGTAACTAGATCT 3.54%

Figure 16: Evidence from previous figures suggests that cleavage occurs on the non-target strand around both the 14th and 19th nucleotides downstream of the PAM. In considering whether this cleavage position might be specific to the guide and matched target pair, this guide suggests that Cpf1 might cleave the same target sequence in multiple positions. It has been suggested that the differences in the length of the overhang, due to cleavage position, may depend on the size of the DNA substrate containing the target site [6] and the length of the spacer sequence [4]. For this data, however, both were the same so there may be further reasons for this variability.

Further Work

As an extension to this work the following questions could be considered:

* Did indels corresponding to a deletion followed by an insertion of a smaller number of bases occur in similar positions to indels that were just deletions?
* To what extent can the cleavage position on either strand be deduced from the sequenced reads without using experimental techniques?

For the first question, these indels corresponded to around 4% of the weighted indel occurrences and if true then the inserted bases may be a consequence of non-homologous end-joining between both sides of the cut with the overhangs “chewed” back. For the second question, if it is possible, can an estimate of the proportion of times cleavage occurred at around the 14th or 18th and 19th nucleotides of the target sequence be made? Although the example of the guide in the last result suggests that cleavage may occur in both positions for the same guide-target sequence, is this actually the case, with cleavage position dependent more on length of guide sequence [4] or size of DNA substrate [6] or do the guide-target sequences themselves have any effect?

# References

|  |  |
| --- | --- |
| [1] | S. Q. T. M. S. P. N. T. N. M. M. W. J. M. L. Z. R. M. M. J. A. J. K. J. Benjamin P. Kleinstiver, “Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells,” *Nature Biotechnology,* 2016. |
| [2] | J. S. G. O. O. A. I. M. S. K. S. P. E. S. E. V. J. J. J. v. d. O. A. R. E. V. K. F. Z. Bernd Zetsche, “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System,” *Cell,* 2015. |
| [3] | J. K. J. K. H. K. W. B. S.-h. Y. &. J.-S. K. Daesik Kim, “Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells,” *Nature Biotechnology,* 2016. |
| [4] | L. S. L. J. Z. X. Z. G. W. J. Lei C, “The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro.,” *Nucleic Acids Res,* 2017. |
| [5] | M. S. J. L. A. V. M. S. J. Y.-M. K. J. W. C. E. W. H. C. K. J.-W. N. &. H. K. Hui K Kim, “In vivo high-throughput profiling of CRISPR-Cpf1 activity,” *Nature Methods,* 2016. |
| [6] | P. A. G. M. Stefano Stella, “Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage,” *Nature,* 2017. |