Alignment_QC

Wes Horton

June 9, 2016

Overview

Based off of the pretty alignments that we have looked at in depth, we believe that a significant proportion of our alignments may be incorrect. This is because we observe multiple times that only 20-25 nucleotides of the V or J sequence aligns to reference, where we would expect many more. When we run the intervening sequence through BLAT, it often aligns to random genes elsewhere in the genome. This is evidence of off-target amplification by our primers.

We hope that some of this may be eliminated by our new PCR conditions, but most likely we will still observe this problem. Another possible way to avoid these sequences in our final analysis is to excise them from our gel prior to sequencing. This would require a significant difference in overall sequence length so that distinct bands will form in the gel. According to DM, a range of 170-240 nucleotides is expected for a proper VDJ sequence and alignment. If we observe alignments that are shorter or longer than this range, they are likely to be the result of off-target amplification by our primers.

There is still the possibility of true alignments existing outside of that range and false alignments within. We can use the alignment length of just the V region as another requirement for "true alignments". We expect false alignments to only align to the primer sequence (22-25 basepairs) and no more. Any sequence who's V alignment is shorter than 30 base pairs is likely to be a false alignment.

Summary of criteria

Alignment length is defined as the first nucleotide of the V alignment to the last nucleotide of the J alignment produced by MiXCR. These values exist in the pretty alignment files, but also in the tab-separated files created by exportAlignments. V alignments are in Best.V.Alignment and J's are in Best.J.Alignment. There are many values in these columns, separated by "|" characters. The beginning of the V alignment is the 4th field in that column and the end of the J alignment is the 5th field in its column. The difference between the two is the alignment length.

1. False alignments have total alignment lengths that are shorter than 170 nucleotides and longer than 240 nucleotides.
2. False alignments have V alignments that are shorter than 30 nucleotides.

Example

First we want to apply our first criterion: false alignments are likely to be outside the range of 170-240 nucleotides

```
##  If total alignment length is outside of this range, it is a potentially off-target alignment
align.outside.range <- align.data[align.data$tot.length < 170 | align.data$tot.length > 240, ]
##  Rename
off.target.1 <- align.outside.range
rm(align.outside.range)
##  What percentage of alignments are considered bad at this point?
percent.bad.1 <- round(length(off.target.1[,1]) / total.alignments * 100, digits = 1)

##  If total alignment length is within this range, it is a potentially correct alignment
```

```
align.in.range <- align.data[align.data$tot.length >= 170 & align.data$tot.length <= 240,]
##  Rename
correct.1 <- align.in.range
rm(align.in.range)
##  What percentage of alignments are considered good at this point?
percent.good.1 <- round(length(correct.1[,1]) / total.alignments * 100, digits = 1)

##  At this point, we have the most general division between "good" and "bad" alignments.
##  Theoretically though, there are most likely false positives as well as false negatives.
##  Is there something that we can define as a correct alignment to use as a checkpoint here?

## Add to summary
summary.df$percent.outside.of.range <- percent.bad.1
summary.df$percent.in.range <- percent.good.1
```

At this point, we have flagged 57% of our reads as "bad", or "off-target", and 43% of our reads as "good". Let's add our second criterion: false alignments are likely to have V alignments less than 30 nucleotides:

```
##  Now we want to extract from the "bad" alignments any that may actually be "good" alignments
##  and add them to the "good"
off.target.2 <- off.target.1[off.target.1$V.length < 30,]

correct.2 <- rbind(correct.1, off.target.1[off.target.1$V.length > 30,])

##  What percent of total alignments are these?
percent.bad.2 <- round(length(off.target.2[,1]) / total.alignments * 100, digits = 1)
percent.good.2 <- round(length(off.target.2[,1]) / total.alignments * 100, digits = 1)

##  Add to summary
summary.df$pct.out.range.short.v <- percent.good.2
summary.df$pct.in.range.or.long.v <- percent.good.2
```

At this point, we have flagged 19% of our reads as "bad", and 19% of our reads as good. Our ultimate goal of this analysis is to determine if there is a significant size difference between good and bad reads that we can utilize during library preparation.

We want to minimize false positives and false negatives. We don't want to have a large portion of our kept reads to be bad alignments, nor do we want to throw out a bunch of good alignments. I think that V alignment length