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. One reason for this suspicion is that 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 to some degree, due to the nature of PCR. 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.

In addition to determining whether or not we can use size-selection in our library prep, we are also interested in characterizing the frequency at which off-target amplification occurs. Finally, we can subset our alignment files to only include alignments that later assemble to clones. It will be informative to look at the frequency of off-target amplification in those alignments as well, to determine if MiXCR is working correctly.

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

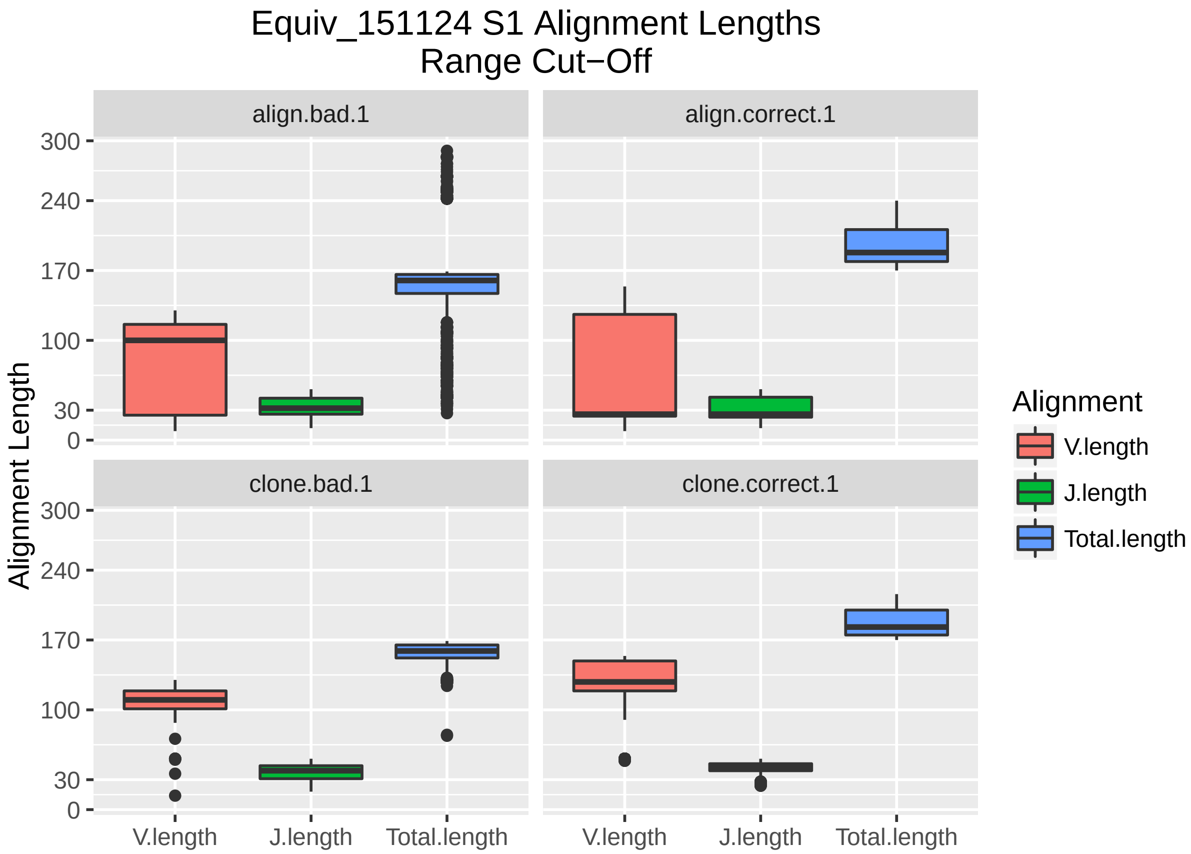
Alignment length is defined as the first nucleotide of the V alignment to the last nucleotide of the J alignment produced by MiXCR. These values exist in the pretty alignment files, but also in the tab-separated files created by exportAlignments. V alignments are in Best.V.Alignment and J’s are in Best.J.Alignment. 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 or longer than 240 nucleotides.
2. False alignments have V alignments that are shorter than 30 nucleotides.

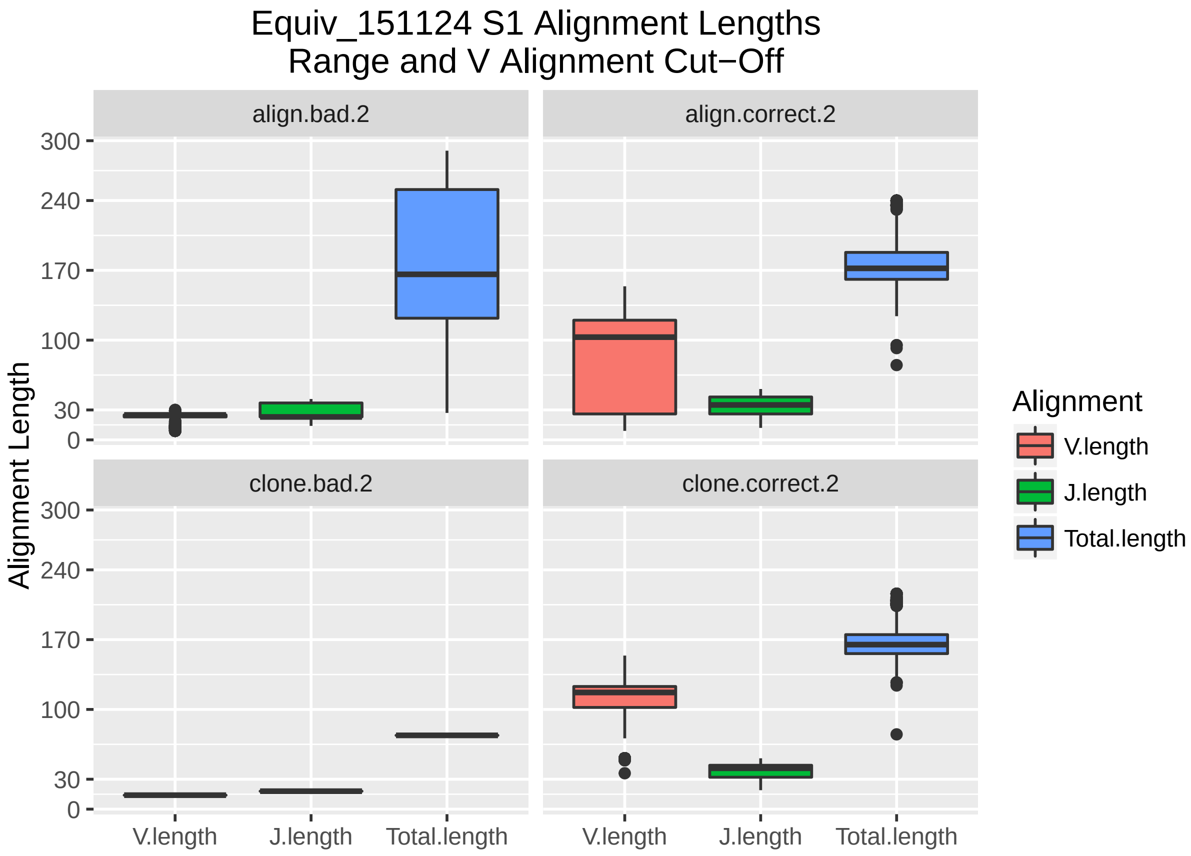
Example

This is a look into one alignment, specifically sample 1 from the equivolume run of DNA151124.

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



At this point, we have flagged 57.2% of our reads as “bad”, or “off-target”, and 42.8% of our reads as “good” in our total alignment. When subsetting by clones, those numbers actually worsen to 67.1% and 32.9% of reads for bad and good alignments, respectively. We can also see, from this plot, that it may be difficult to use a size cut off. The difference between total alignment length for good and bad alignments is not very large and also seems to be an artifact of V alignment length. Let’s add our second criterion: false alignments are likely to have V alignments less than 30 nucleotides:

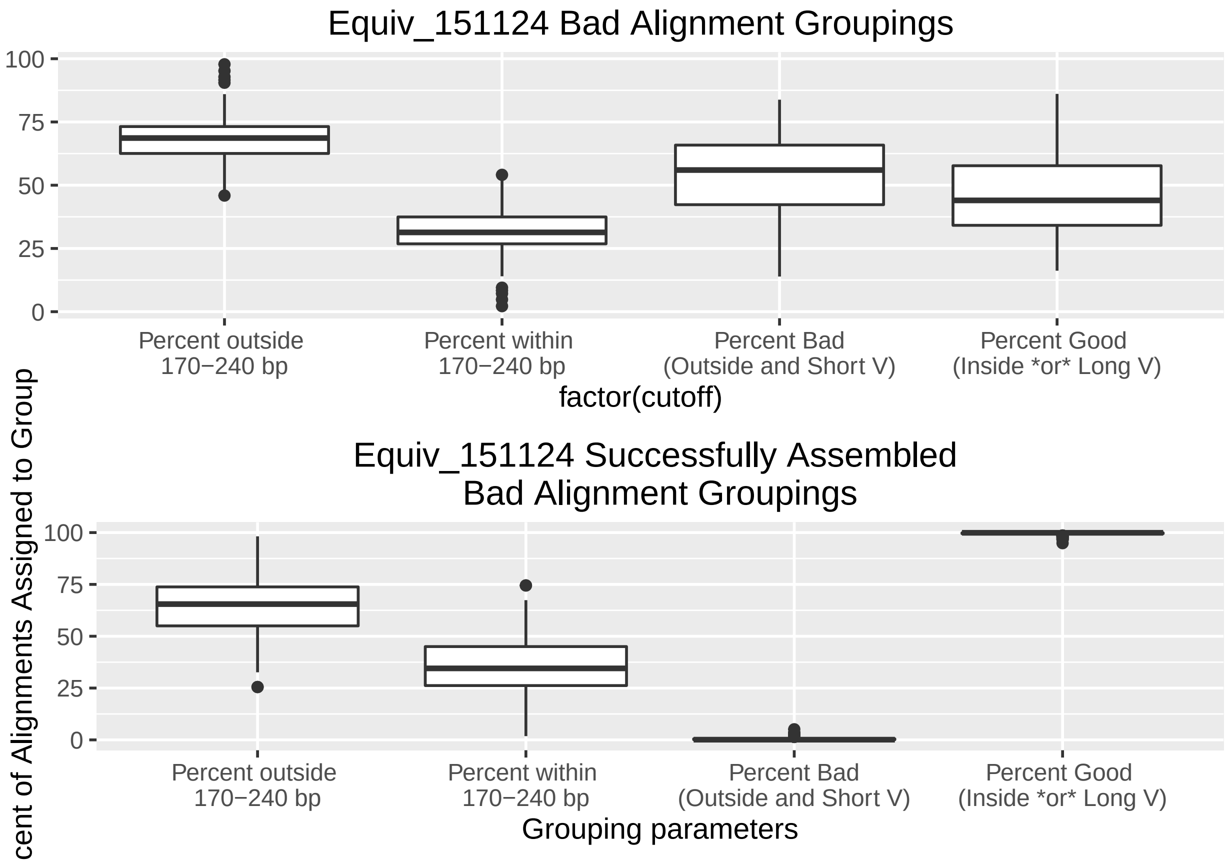


At this point, we have flagged 20.6% of our reads as "bad", and 79.4% of our reads as good. In comparison, looking at only our alignments that successfully assemble to clones, we've flagged 0% of our reads as bad and 100% of our reads as good (Although there is actually 1 read that is still considered bad, but I rounded the percents to 2 decimal places).

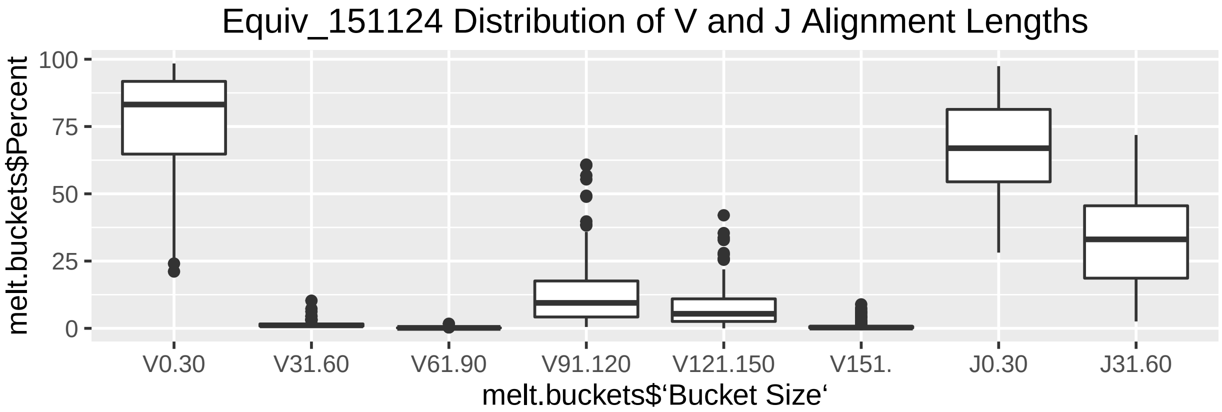
Our ultimate goal of this analysis is to determine if there is a significant size difference between good and bad reads that we can utilize during library preparation. From these results, it looks like we may not be able to use a size cut-off during libarary preparation. Our "bad" reads dropped from an initial 57.2% of reads to a final 20.6% of reads for our total alignments, and we drop from 67.1% of reads to 0% when using successfully assembled reads, an essentially perfect filter. A large percentage of reads are outside of our defined range of 170-240, but many of those are rescued by identifying their V alignment length.

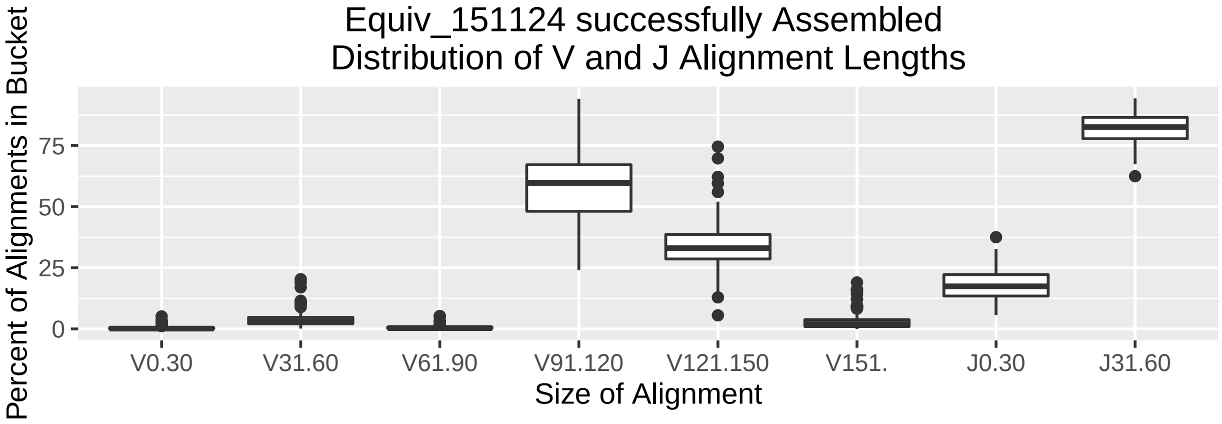
Before we think about alternative parameters to use as cut offs, let's look at an overall summary of the entire equivolume 151124 batch. We want to look at the above information, as well as information specific to V and J alignments

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```





The first set of plots shows us that Sample 1 is similar to other samples. We rescue a good portion of the reads when we take V alignment length into consideration (Change between cols 1 and 3 or 2 and 4).

From the second set of plots, we can see that there is a clear divide between V alignments, but not so clear for J alignments, when looking at all alignments. There were no J alignments greater than 60 base-pairs, so I did not include them in the plot. When looking at just alignments that successfully assemble, we see a similar distribution of V alignment lenghts, except we have very few alignments less than 30 bp, as we would expect. These plots suggest that using V alignment length as a cut off is appropriate, but that we should not use J alignment length in the same fashion.

Taken together, it may not be advisable to use a size selection step during library prep because many good alignments have a size that is outside of our desired range.

Moving Forward

1. I don’t think we should implement a size selection step, but we should see if the new data follow these trends before making a final decision

2. MiXCR seems to be doing a good job with the correct sequences. When we filter based on what we expect to see, the percentage of assembled reads (from total aligned reads) approaches 100%. That being said, additional quality metrics may be prudent.

   - V/J and CDR3 identification comparison with MIGEC
   - BWA-MEM aligner comparison