MiXCR QC Analysis

Wes Horton

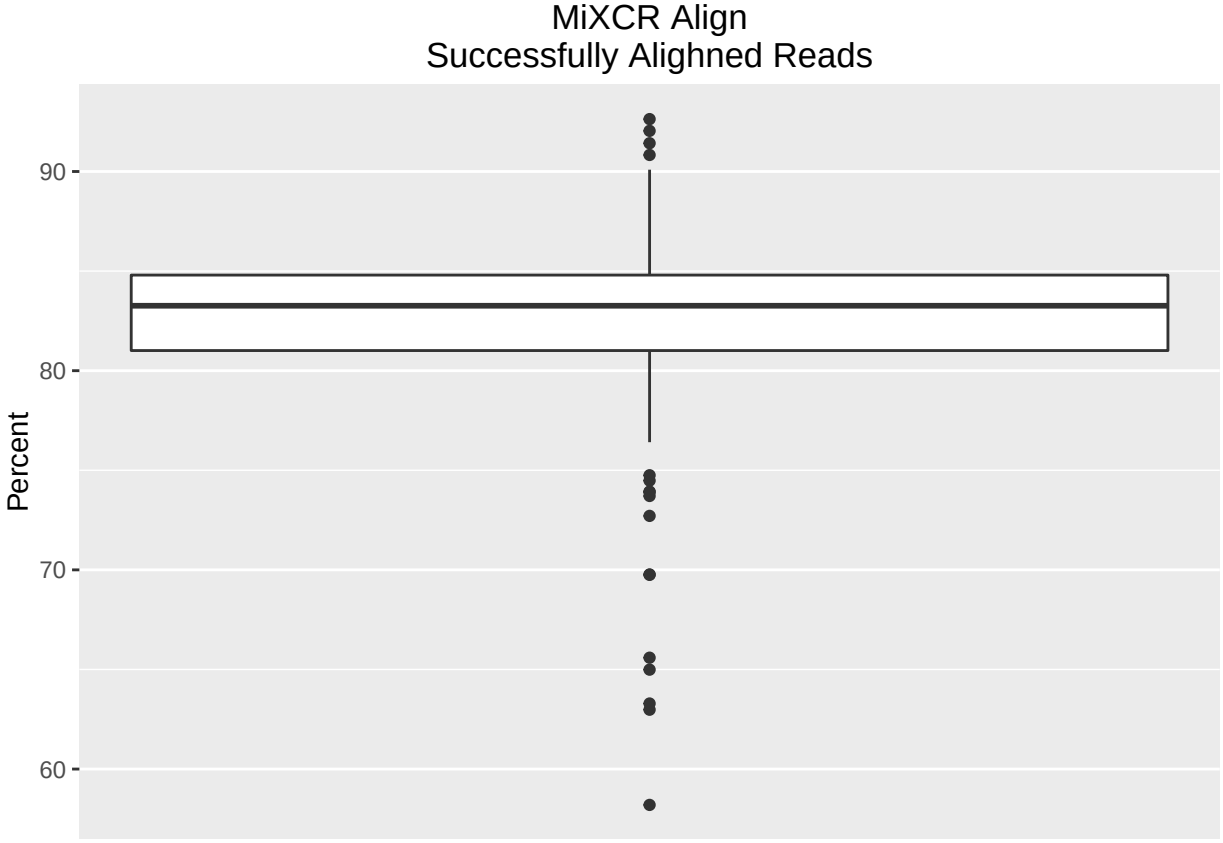
May 6, 2016

Summary

We need to determine if we can use MiXCR clonotype count outputs as a proxy for depth of coverage. To do so, we need to figure out how reads are aligned and assembled. Are they grouped together too often, what is the reasoning behind the grouping, what happens if we change certain parameters, etc.

Alignment

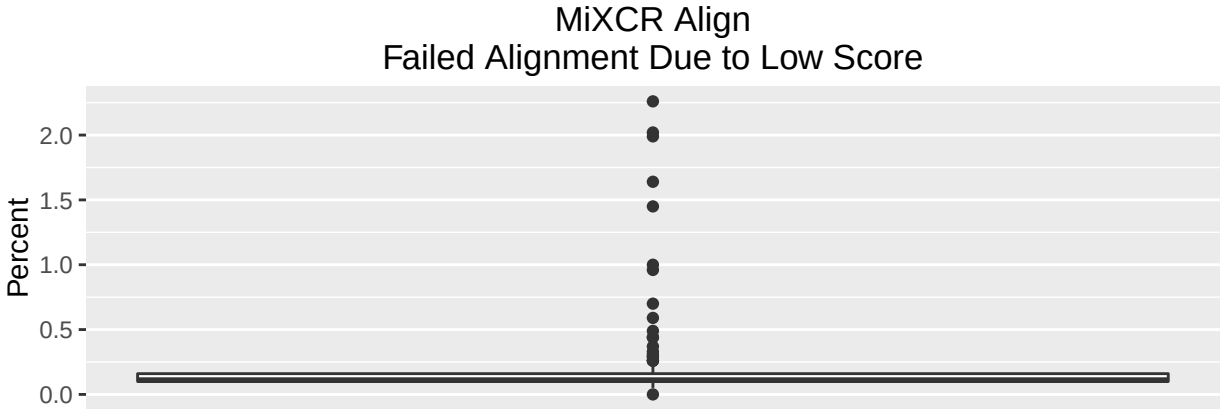
First, lets look at a boxplot of the percentages of aligned reads for each sample as well as a summary of the distribution:




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   58.20   81.01   83.26   82.39   84.80   92.63
```

We see that most of the samples aligned greater than 80% of their reads, but a few have relatively poor alignments. Why is this?

```
## Loading required package: grid
```




```
## [1] "Failed Due to Low Score:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1000  0.1100  0.2017  0.1600  2.2600
```

```
## [1] "Failed Due to No J Hit:"
```

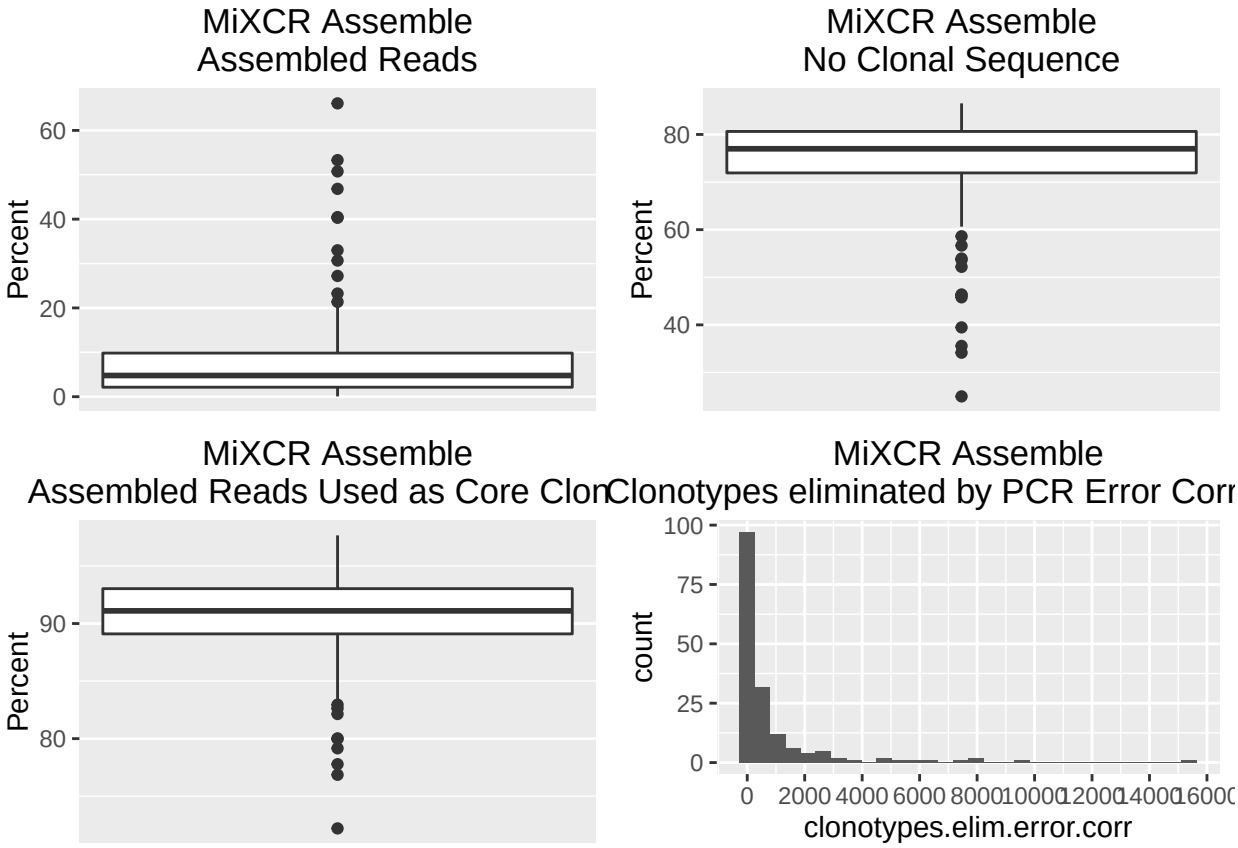
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.33   15.07   16.64   17.41   18.86   41.66
```

Looks like most reads are not aligning due to a lack of J hit. Where are these reads coming from? How do 20% of our reads not have a matching J alignment? Should we relax the parameters for calling a hit? We could potentially extract these reads from the fastq file (I think) and re-run just them through mixcr with relaxed parameters and see how many more we catch.

Assemble

Lets do the same for the assemble QC file.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



These data don't look very good. A majority of the samples assemble less than 20% of their reads to clonotypes. We also see that most unassembled reads were not assembled due to a lack of clonal sequence (top right boxplot). The other reasons are due to low quality (none were dropped) and due to failure to map to a core clone:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.0900  0.1600  0.2644  0.3200  1.5700
```

Which is also pretty small. From the lower left plot, we see that of the reads assembled, most of them are used as core clonotypes. Finally, from the histogram, we see that quite a few clonotypes are eliminated from the overall count due to the PCR error correction, although many lose fewer than 700. Let's look at the summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    24.0   177.0   876.7   625.0 15380.0
```

Compare clonotypes to Reads

To Do: Not sure if this is appropriate or not.

Let's make a few comparisons here. Total clonotype count of a sample can be compared to the reads used in assembly, which should give us a handle on how many reads we're losing to clustering? We can also compare the distributions of clonotype counts for each sample.

```
clone.dir <- "~/Desktop/OHSU/tcr_spike/data/equiv_DNA160107LC/clones/"
clone.files <- list.files(clone.dir)
```

```
clone.files <- clone.files[order(as.numeric(gsub(".*_S|_align.*", '', clone.files)))]

for (i in 1:length(clone.files)){
  curr.clone <- read.delim(file.path(clone.dir, clone.files[i]), header = T, sep = '\t')
  clone.count <- sum(curr.clone$Clone.count)


}
```