160609 MiXCR QC

Wes Horton

June 22, 2016

Overview

DNA160609LC is the most recent batch of samples, received 6/17/16. There are many different treatment groups, as outlined in sample_identification.csv, found here https://ohsu.app.box.com/files/0/f/8485211449/DNA160609LC. We have monoclonal (El4, Ot1, P14) and wild-type spleen samples, wild-type blood and tumor samples, and epithelial samples for negative controls (EpH4 and D2OR). Samples either received synthetic template spike-ins prior to PCR, or did not. A subset of samples received spike-ins and no genomic DNA. We also have samples that are just water and spike or no-spike. Need to determine if there is a difference between the only spike-in samples and the water+spikes samples.

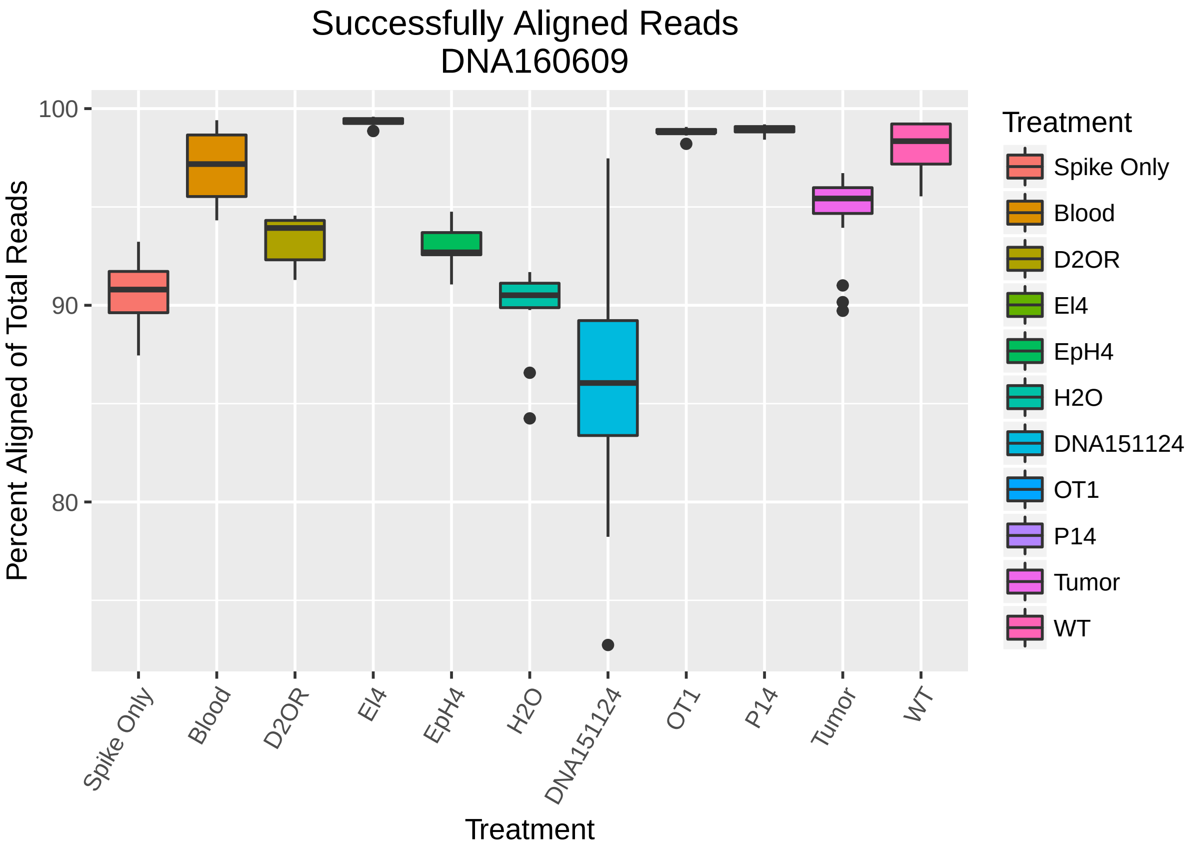
Objective

Ultimately, we want to determine if we can use MiXCR clonotype count outputs as a proxy for depth of coverage. Before we can do that, however, we must confirm the accuracy of the clonotype identification that MiXCR performs. Specifically, we want to know the percentages of reads aligned and assembled, and try and get some grasp on the accuracy of those alignments and assemblies. We will use DNA151124LC as a comparsion batch.

Results

Alignment

From previous analyses, we expect around at least 80% of our reads to align. Most failed alignmets have been due to bad J alignments. Let's see if this holds true.



We see that all of our treatments perform better than the DNA151124 group. A few observations:

1. The three monoclonal sets have very high percentages, with a small range, which is expected.
2. Many of the spike-only reads align, which they shouldn't for two reasons.

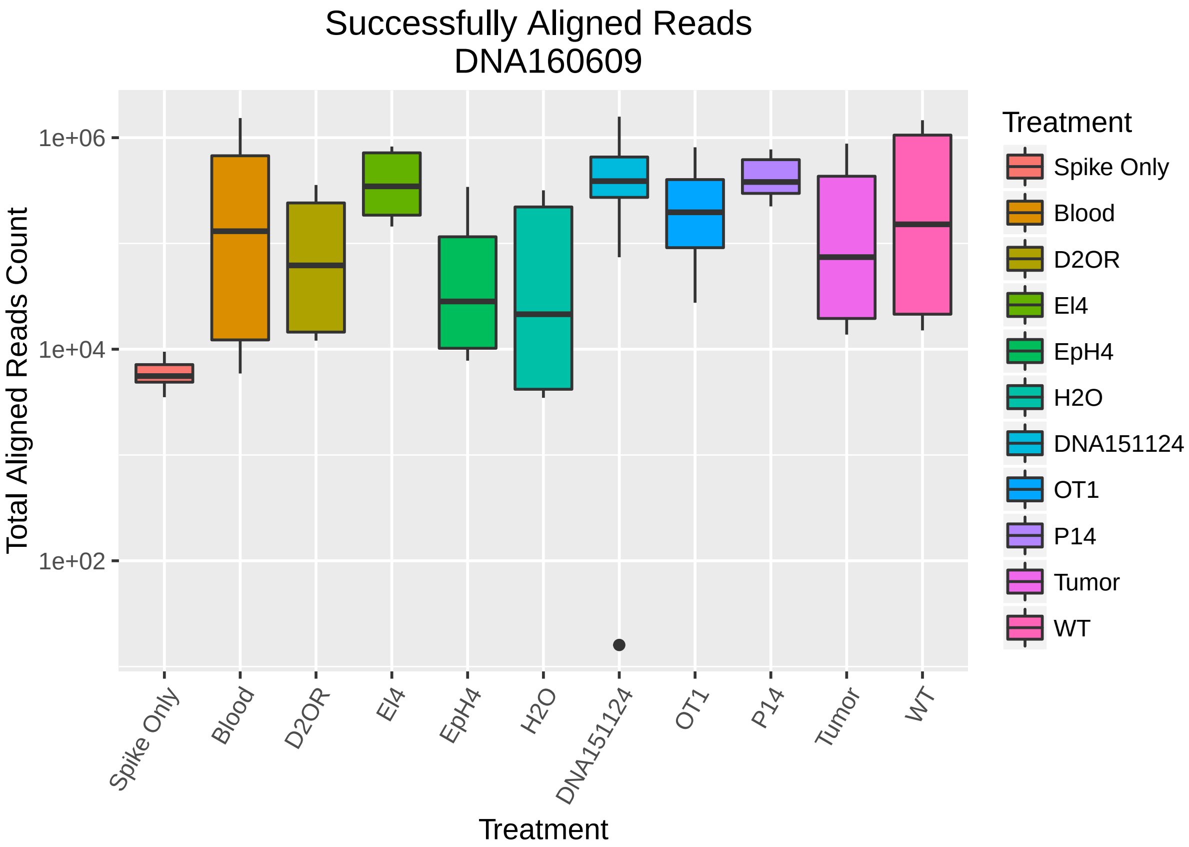
   A) They should have all been removed in the remove.spikes step
   B) Those that we missed only have primer sequences plus a little more

3. Many of the H2O reads align as well.

   A) Approximately the same amount as the spike-only. B)What is going on here that produces so many reads that align?

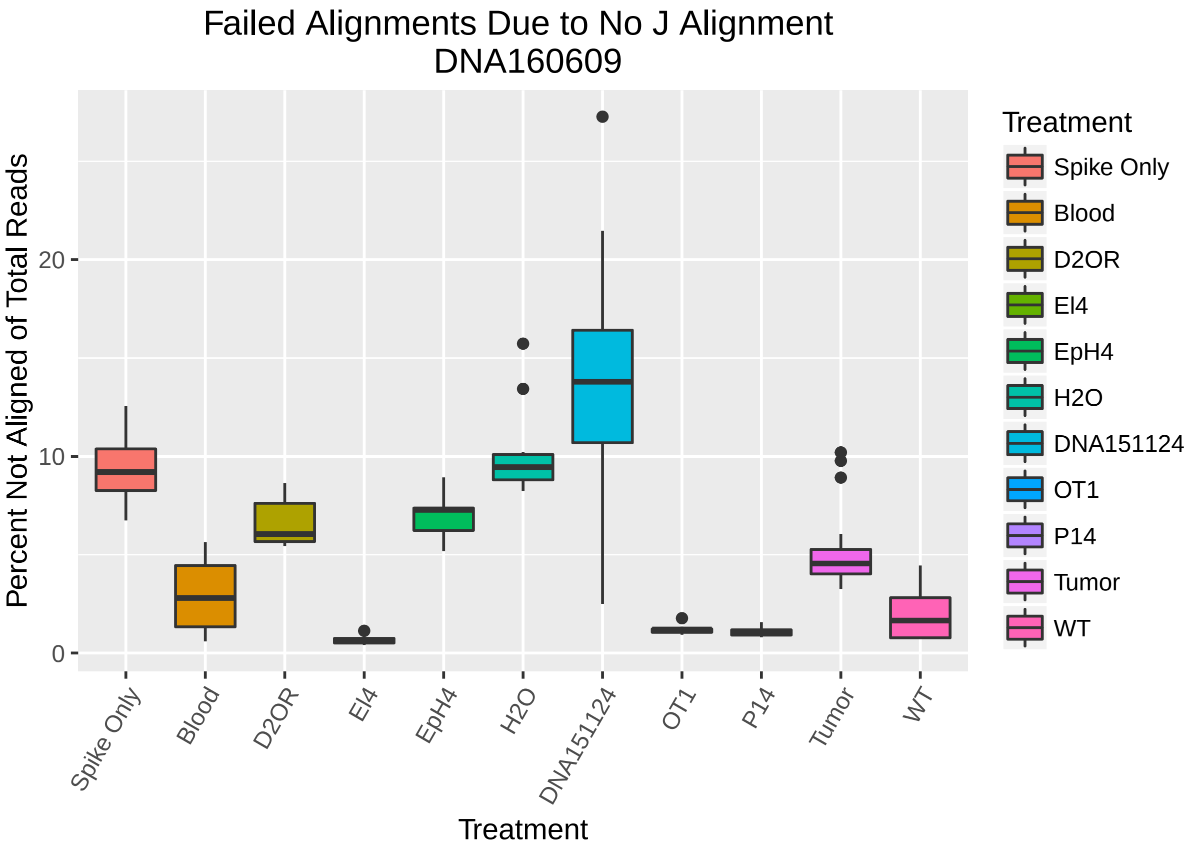
4. Many epithelial reads align as well (EpH4 and D2OR)

We'll have to pay attention to how many reads of these different groups assemble. We can also check their V alignment lengths and see if just the primers are aligning, that should back up whatever we find in the assembly information. One other thing to look at is the absolute amount of reads that align. These negative controls should have significantly fewer reads.



From the total alignment counts, we see that the spike only have much fewer reads, which is reassuring. The epithelial samples and water samples have lower as well, but not as low and not too much lower than the tumor samples. Still need to figure out how we're getting so many reads from these sample types.

A glance at the alignment QC table suggests that again the major reason for failed reads is lack of J hits. The maximum percentage of failed alignments due to low score is 0.1% for the entire 160609 batch and 0.31% for 151124. Let's look at it's distribution:



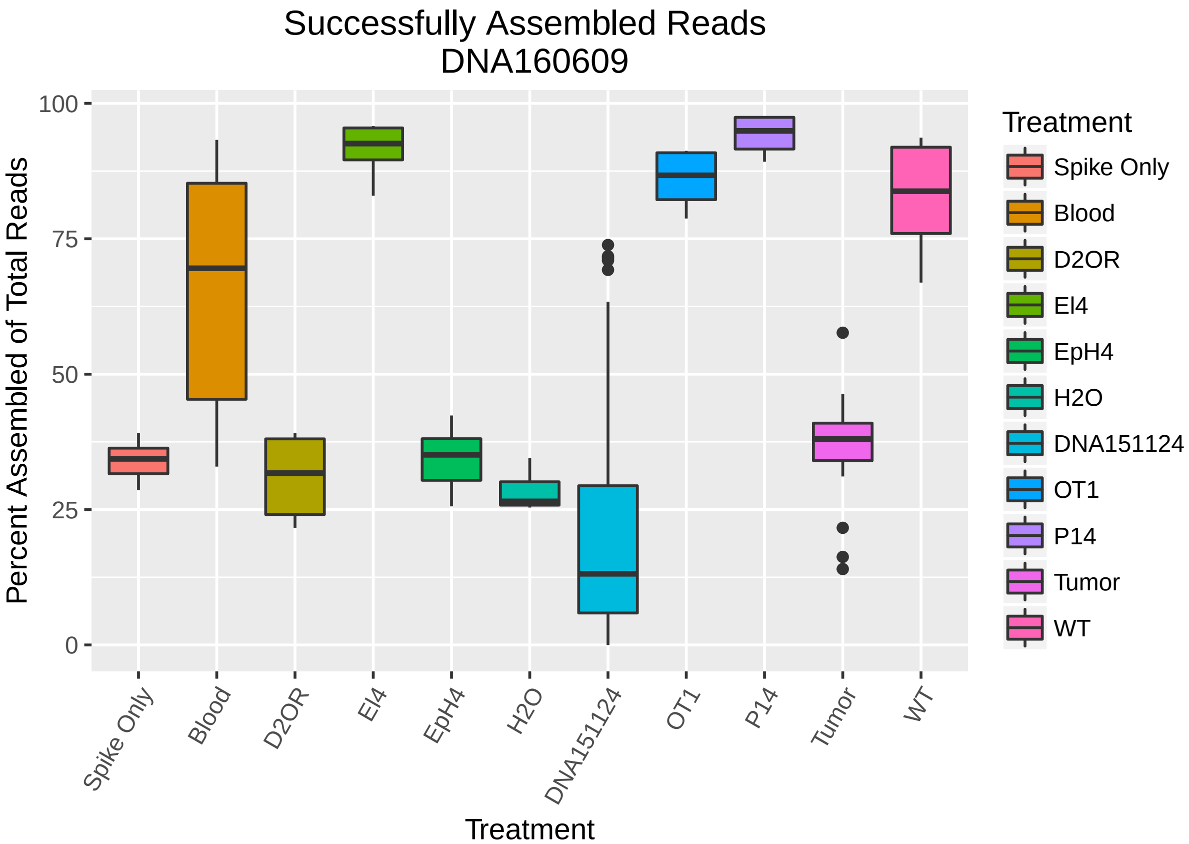
This doesn't really tell us much. It's essentially the reverse of the previous plot.

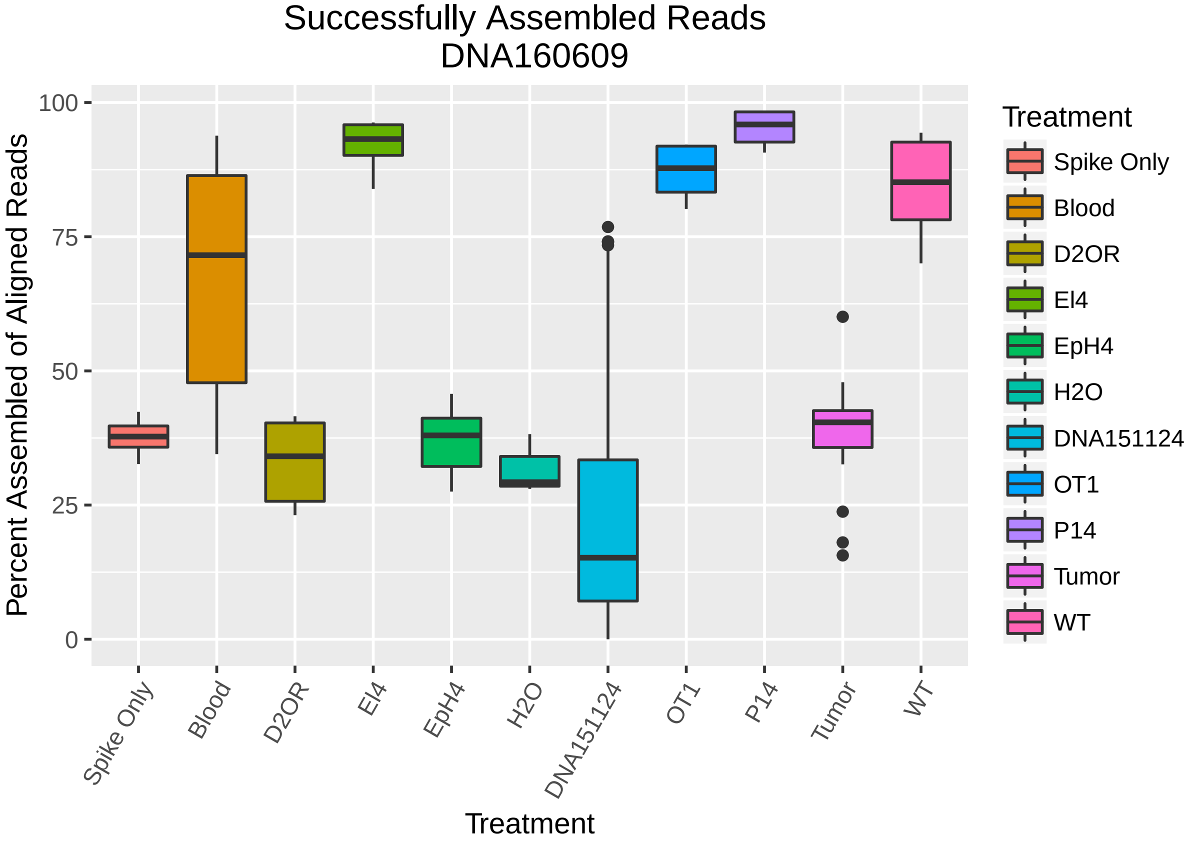
Alignment Conclusions

As a reminder, the alignment step attempts to align portions of our reads to germinle V, D, and J sequences, although only V and J identity are considered when determining whether or not to keep an alignment or not. In general, these results are promising and our alignment percentages are better than in previous batches. We need to keep in mind the high alignment percentages of the spike only and epithelial samples as we move forward.

Assemble

From our previous analysis, we determined that around 15-20% of our reads successfully assembled to final clones. We hope that the improved PCR conditions have increased these percentages.

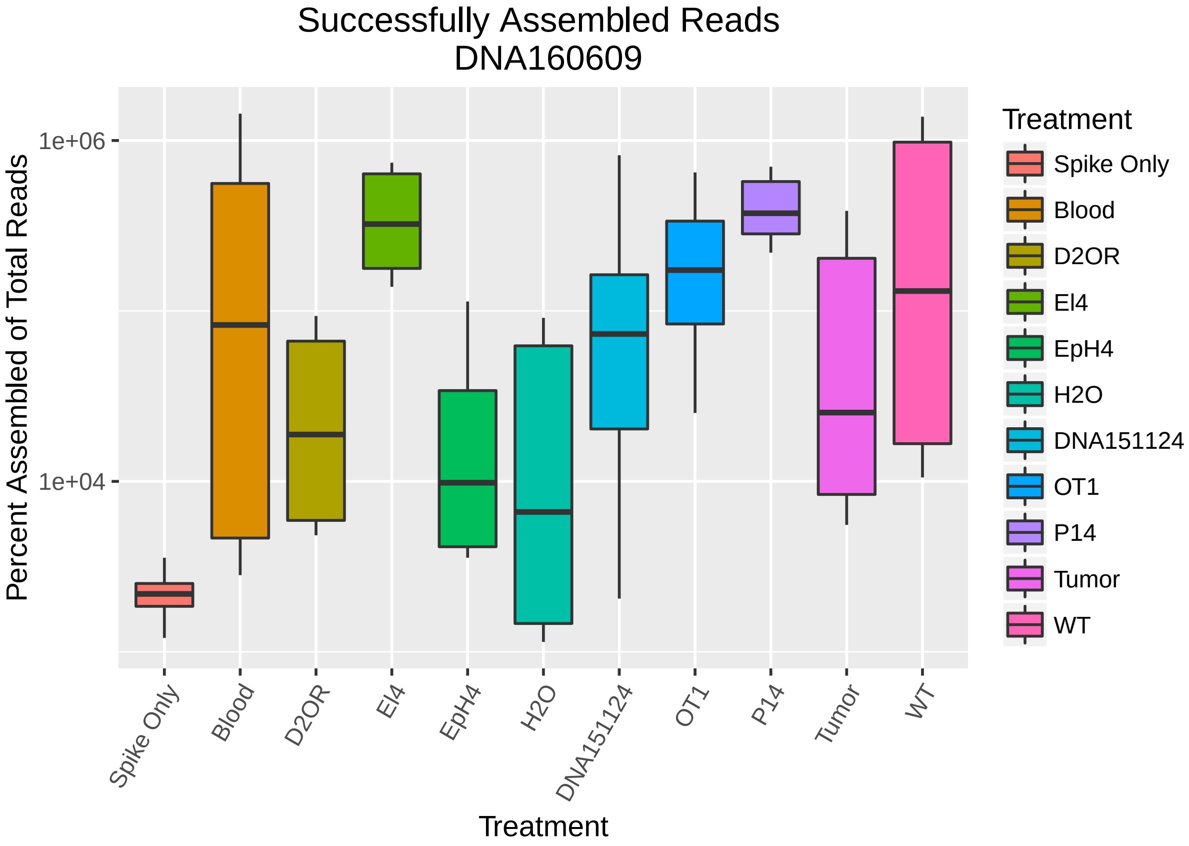




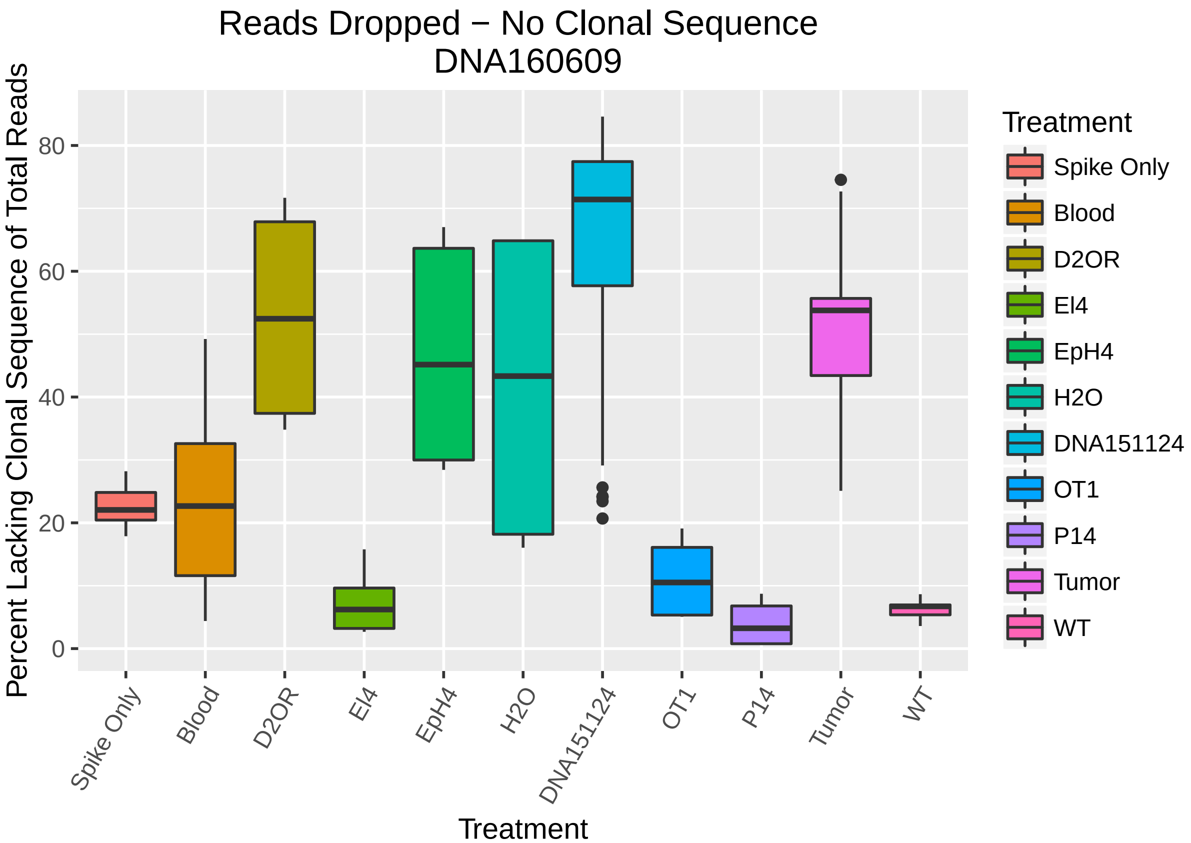
These results are promising as well. My observations:

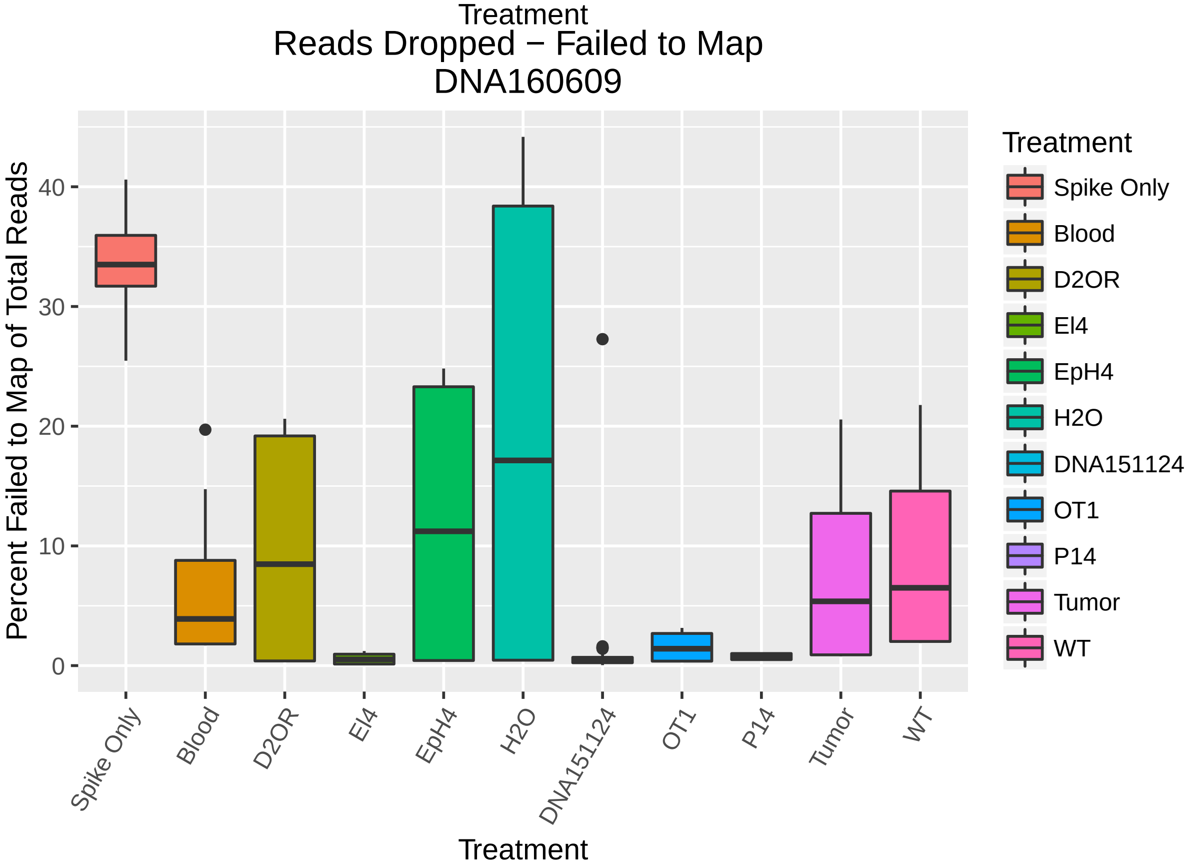
1. Again we see consistently high assembly percentages for the monoclonal mice, as we would expect.
2. All of our negative controls have the lowest percentages assembling, but are similar to the tumor samples.

Again we can take a look at absolute counts:



Not as informative due to the broad ranges, but again we see similar distributions for our negative controls and the tumor group. We can also look at various reasons for failing to assemble:





All of the “positive controls”, i.e. the spleen samples, have very low percentages of missing clonal sequence reads, which is good to see. Again the epithelial cells and H2O controls are similar to the tumor samples. All-in-all, everything is better than the previous batch.

Assemble Conclusions

Marked improvement over previous batches. We need to figure out what's going on with the epithelial and H2O samples that are causing so many reads to be assembled. Can take a look at V alignments to maybe see if they're just aligning/assembling to primers. Spike only reads are also curious. It's good to see that their total numbers are much lower than the other samples, but we still shouldn't see any if our spike removal is good. One possibility is that instead of primer-dimers, primers are amplifying each other further and don't get excised in the size-selection step.

Moving Forward

I still need to do the alignment QC, which could help us determine the quality of the alignments and assemblies. After that, taking a look at some pretty alignments of the epithelial and water samples could be informative as well.