Across Sample Aggregation of Spike Percentages

Wes Horton, Burcu Gurun-Demir

April 25, 2016

Distribution of Spike Percents in a batch

In a given batch, we have approximately 170 samples (give or take a few). Each sample has a spike percentage that is calculated as the number of spiked reads divided by the total number of reads. All three of these values are calculated in the count.spikes QC script for the 9-bp spike.

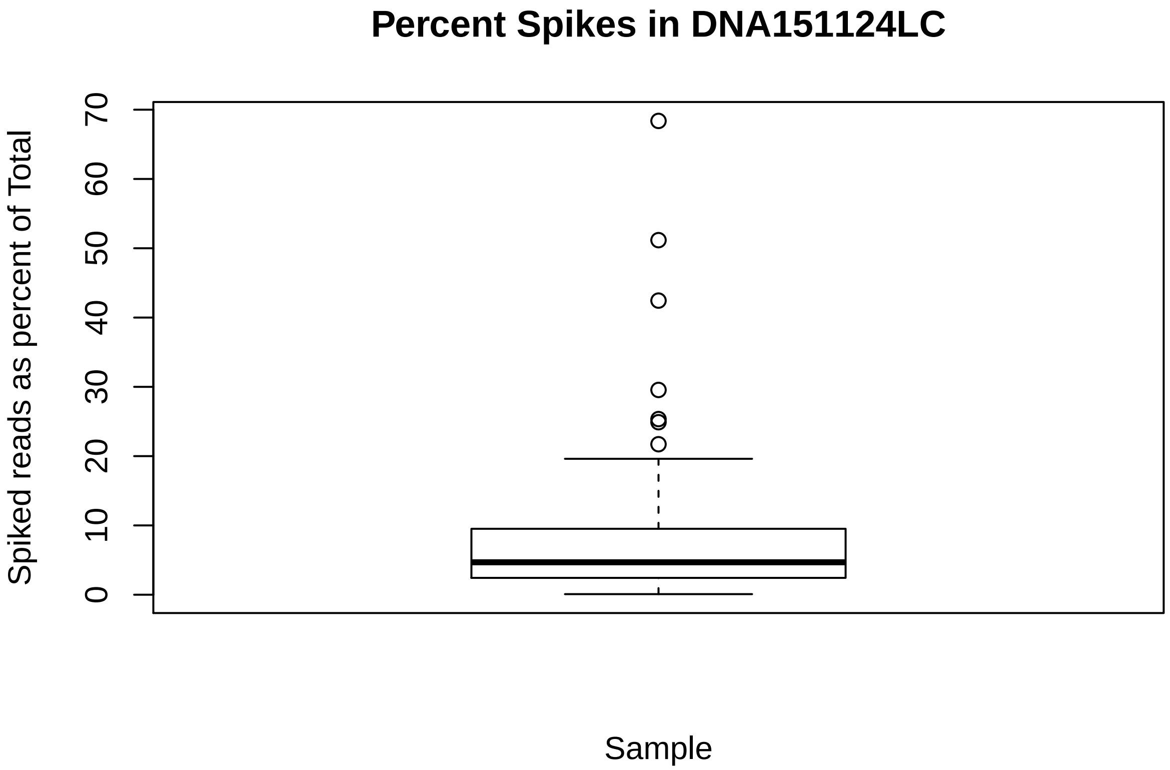
Set up

Before we begin, we must first read in our data. We want to grab the aggregate 9-bp spike count qc file as well as the metadata file for this particular batch. The qc file contains one row for each sample, with a variety of columns. Those of interest columns 2, 3 and 4. They correspond to total reads in the file, number of spiked reads, and spiked reads as a percent of total reads, respectively. The metadata file will be used later for when we subset by sample type. It contains one row for each sample, and specifies sample type and treatment.

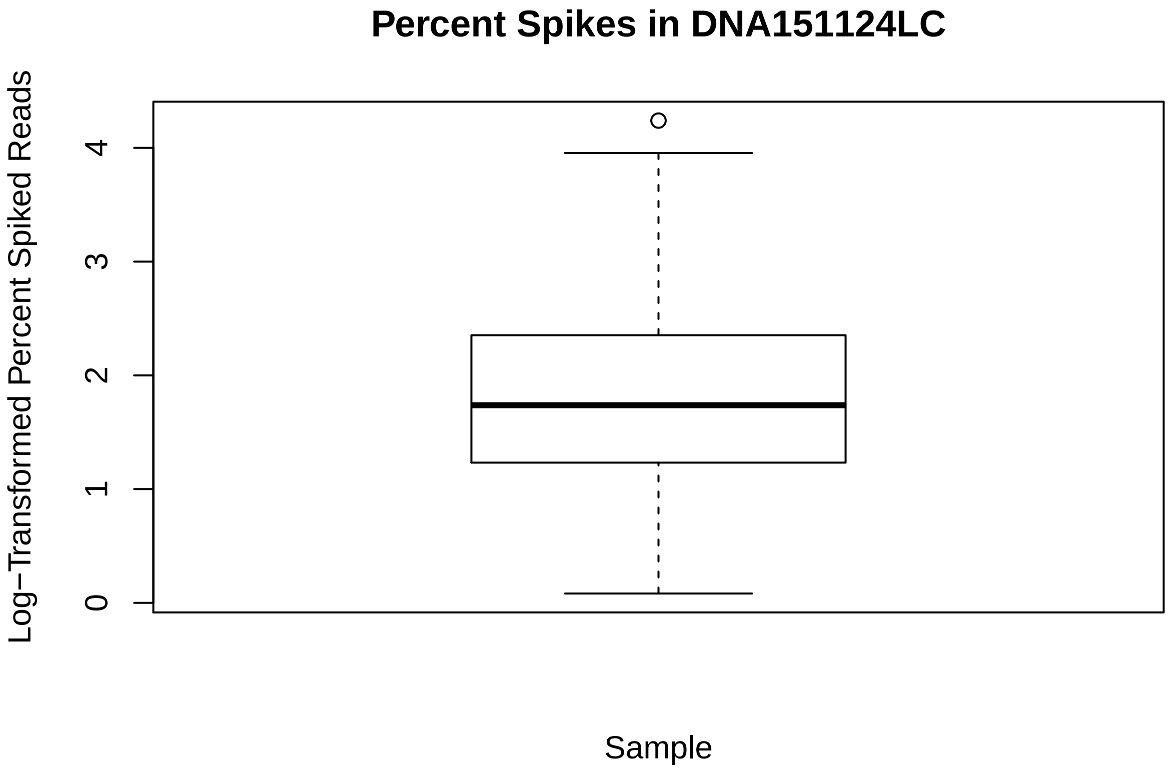
```r
# Read in qc file
qc.file <- "~/Desktop/OHSU/tcr_spike/data/DNA151124LC/QC/9bp.count.spikes.QC.summary.txt"
qc.data <- read.table(qc.file, header = T, sep = ',', stringsAsFactors = F)
# Clean up table
# Remove file path from sample ID
new.ids <- strsplit(qc.data$sample.id, split = "/")
new.ids <- sapply(new.ids, function(x) x[12])
qc.data$sample.id <- new.ids
# Sort by sample number
qc.data$num <- as.numeric(gsub(".*_S|\\..*", '',qc.data$sample.id))
qc.data <- arrange(qc.data, num)
# Remove Sample 142 because it has erroneous data
qc.data <- qc.data[-142,]

# Read in metadata file
metadata.file <- "~/Desktop/OHSU/tcr_spike/data/vj_metadata/151124_qc_metadata.txt"
metadata <- read.table(metadata.file)
```

Overall Summary

First, we'll produce a boxplot of all of the samples in the batch, so that we can get an idea of the distribution and identify any outliers we may have.



We can see that there are quite a few small percents, and a few very large ones. In order to get a clearer picture of these small values, we will now present the same figure with spike percents transfromed into log(percent).



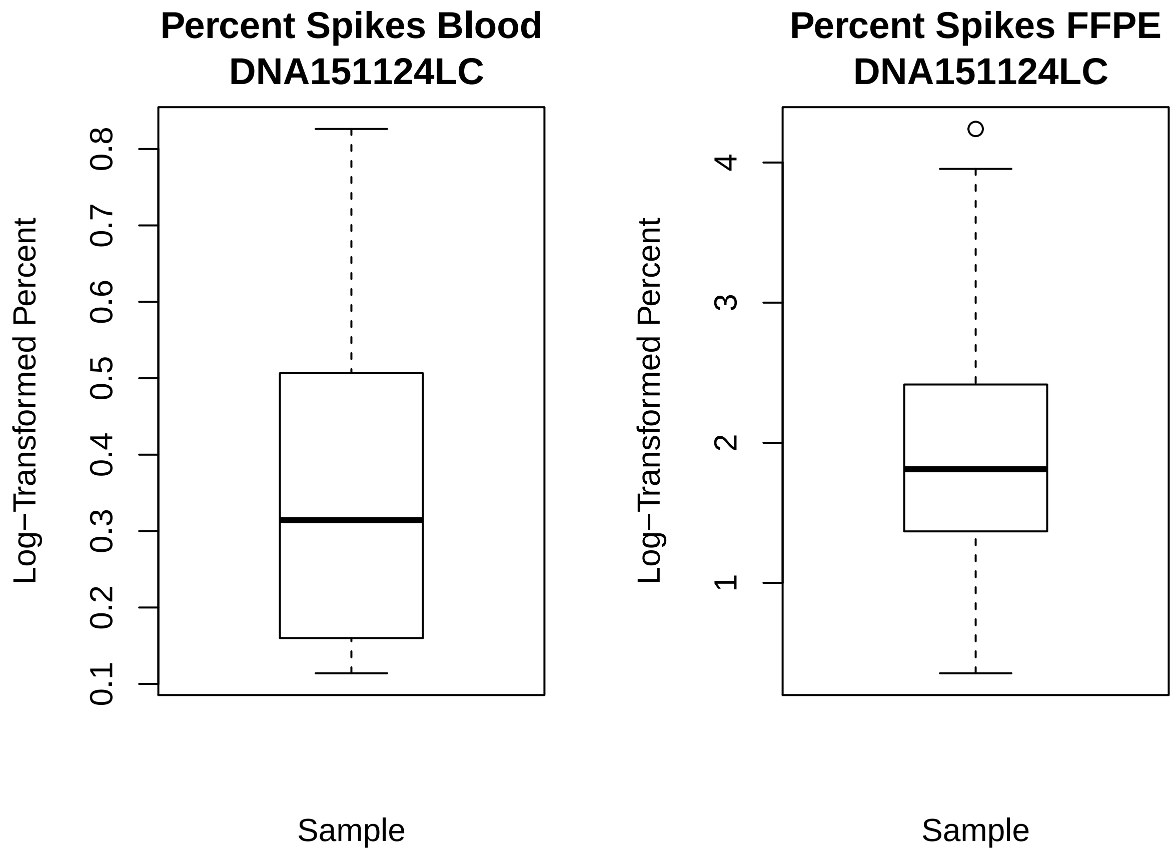
Summary by Sample Type

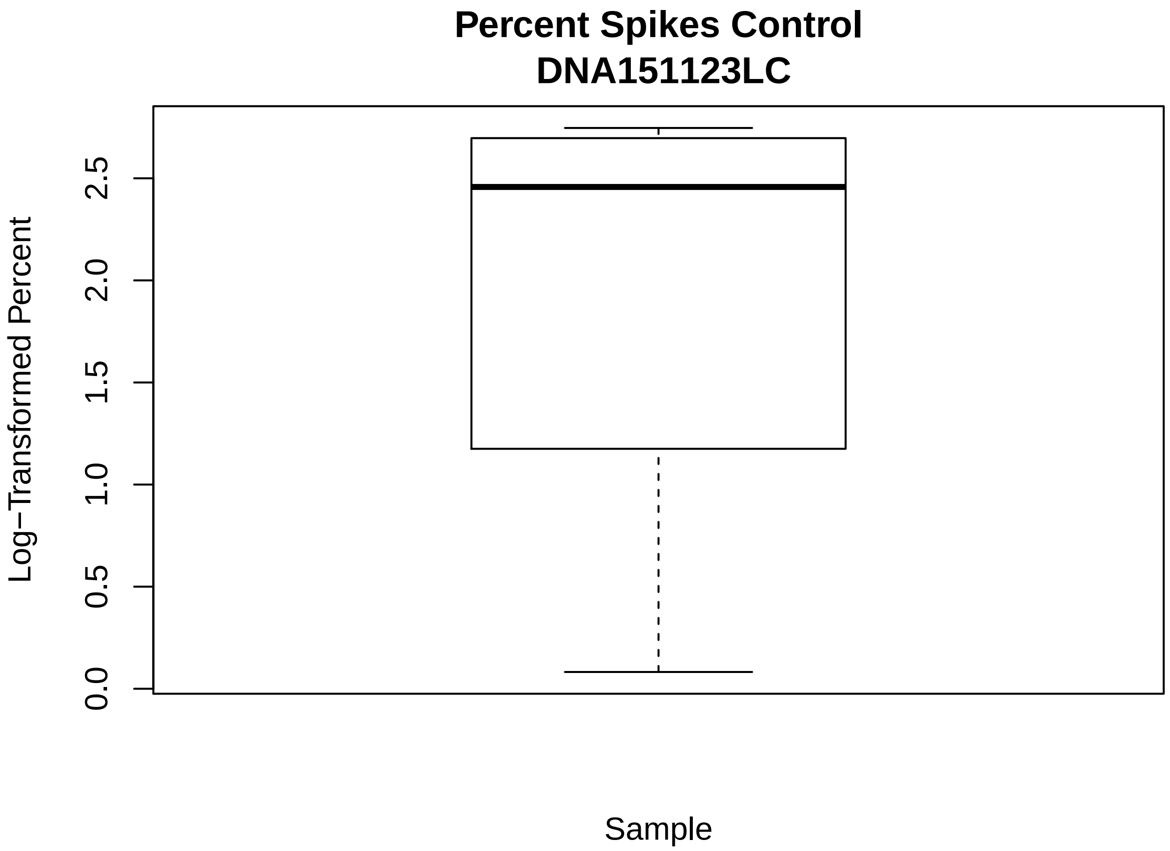
The above plots may not be very illuminating due to the inclusion of different sample types within this batch.

```
summary(metadata)
```

```
##                                        V1                   V2
##  DNA151124LC_1_S1.assembled.qc.txt    :  1   gembtkiIgG2b:16
##  DNA151124LC_10_S10.assembled.qc.txt  :  1   gemIgG2b    :16
##  DNA151124LC_100_S100.assembled.qc.txt:  1   aCD4        :10
##  DNA151124LC_101_S101.assembled.qc.txt:  1   aCD8        :10
##  DNA151124LC_102_S102.assembled.qc.txt:  1   gemaCD4     :10
##  DNA151124LC_103_S103.assembled.qc.txt:  1   gemaCD8     :10
##  (Other)                              :163   (Other)     :97
##        V3
##  blood  : 16
##  control:  4
##  ffpe   :149
##
##
##
##
```

We can see that a majority of the samples are ffpe tumor, but we have a few blood samples as well as some control. TO DO: what are these controls?
Let's subset by sample type and create one boxplot for each.





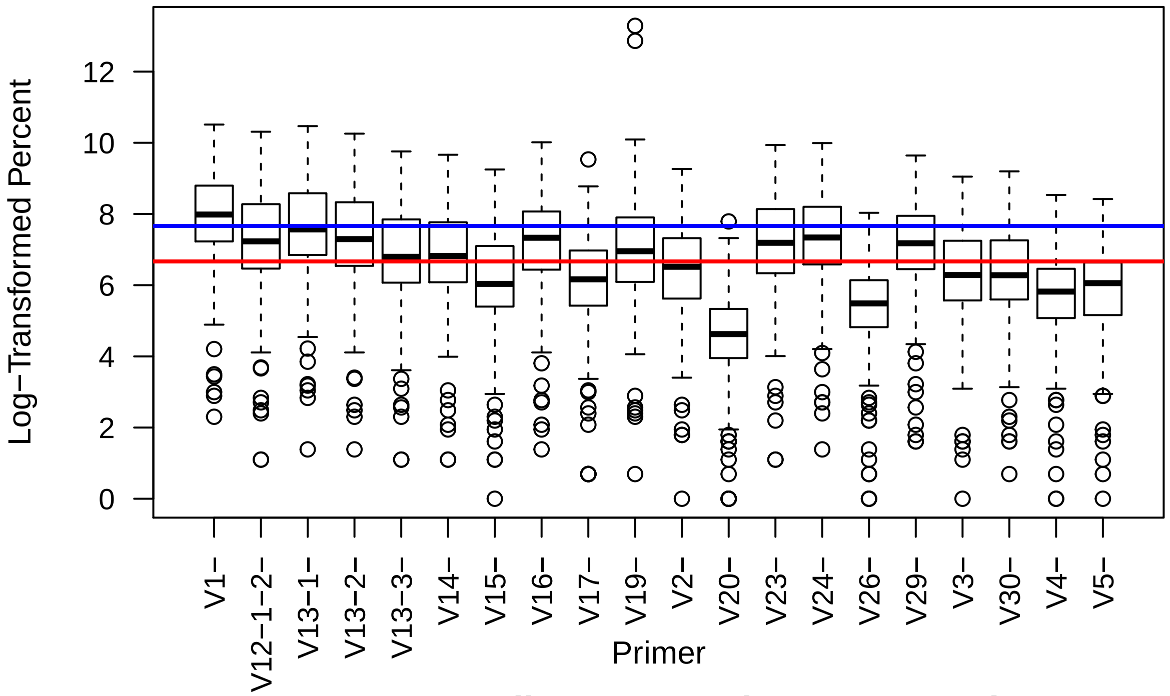
We see that the blood data has a distribution skewed towards lower values and that FFPE looks slightly more normally distributed. Not sure what else to take from these boxplots.

Summary of Forward and Reverse Primers

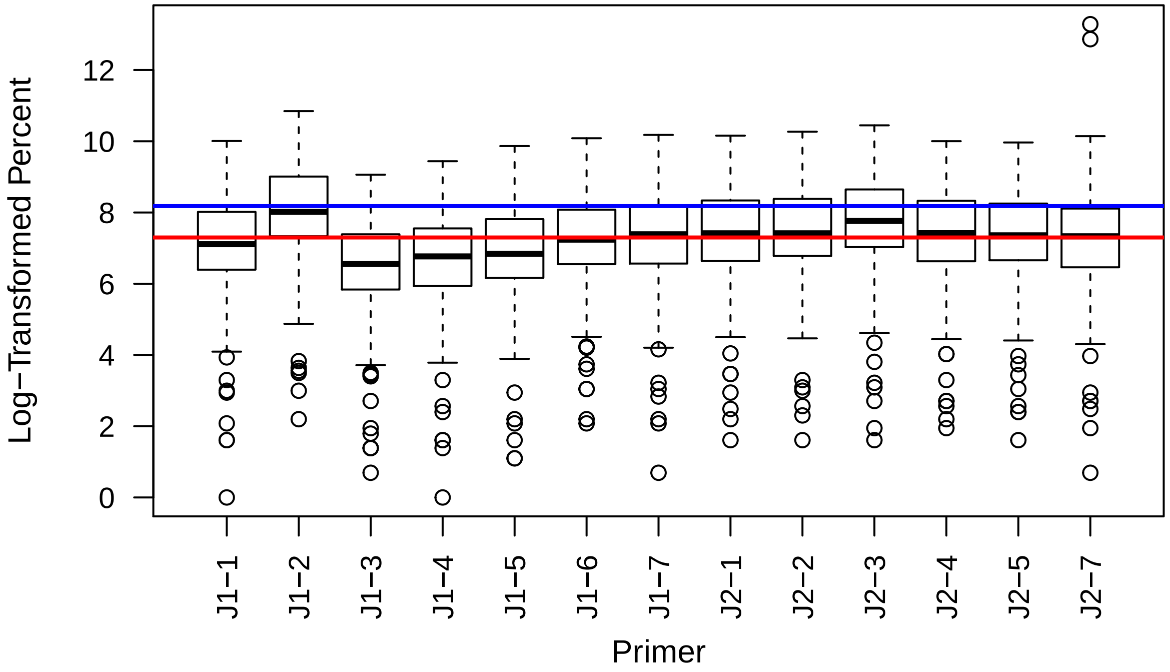
We have 260 synthetic templates that each correspond to a unique forward and reverse primer pair. These 260 primer pairs are comprised of 20 forward (bind to V region of CDR3) and 13 reverse (bind to J region of CDR3). It will be informative to look at the total spike reads (as a percent of total reads) for each of the V primers and for each of the J primers.

First let's look at the V primers. We need to group the spike counts by each V primer, i.e. sum the counts for V1J1, V1J2, V1J3, etc. for all twenty V primers. To do this, we need to use the 25-bp qc file instead of the 9-bp file. We also need to read in the original spike file so that we can convert the DM_# labels into V/J labels

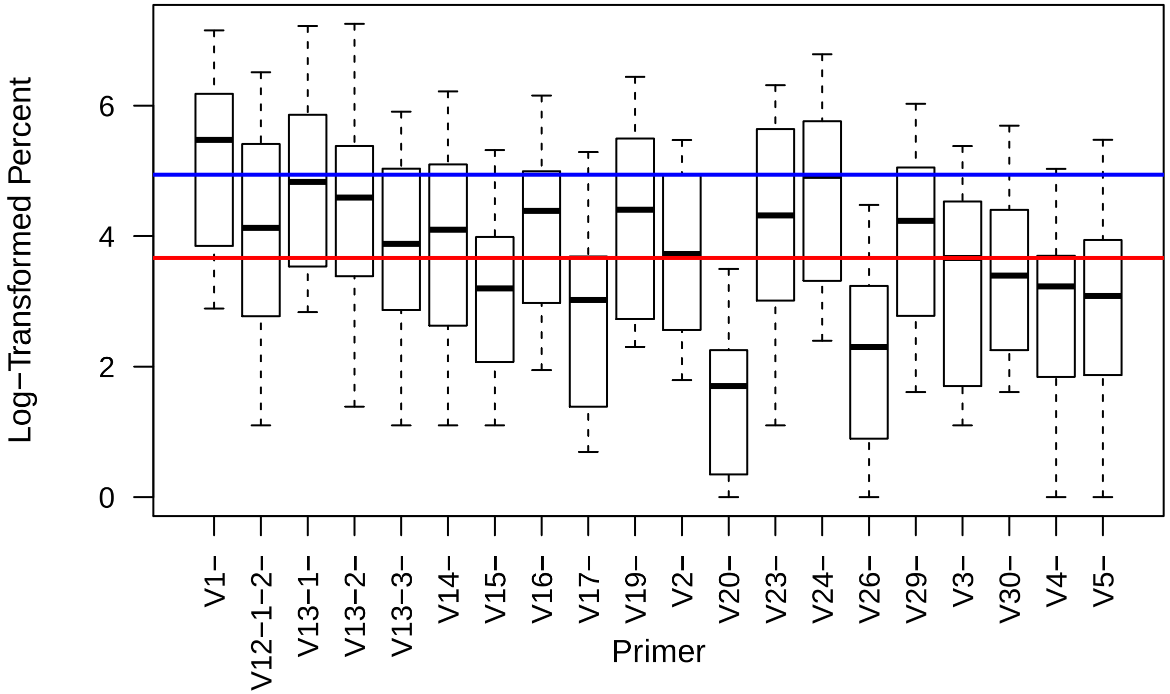
DNA151124 Spike Percents by Forward Primer



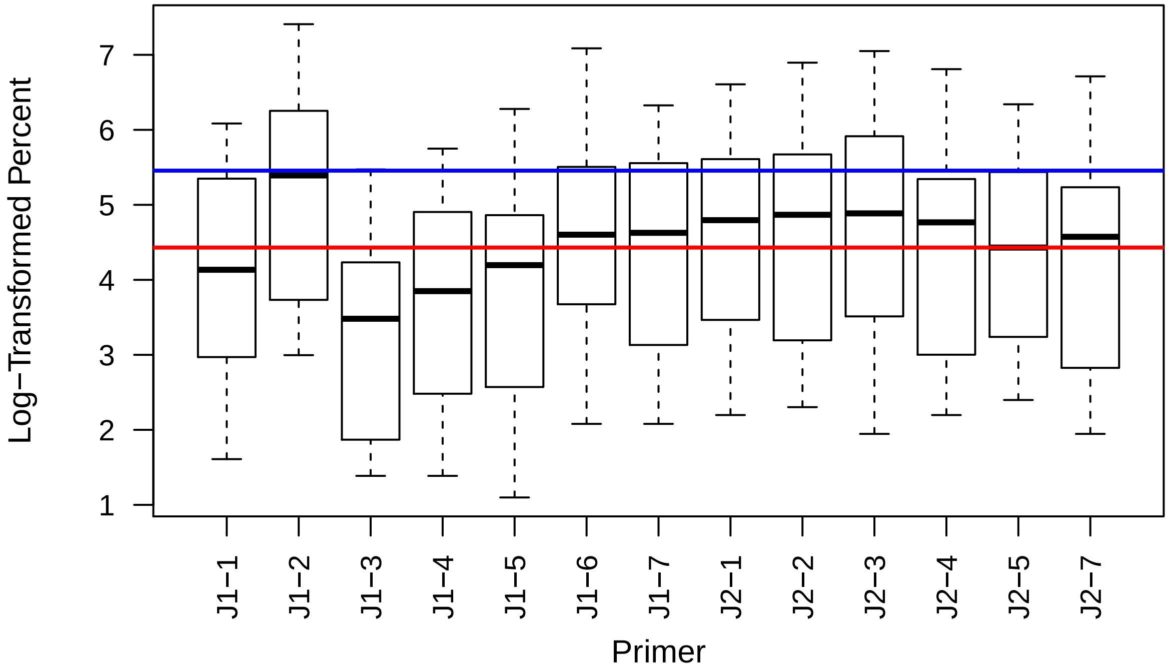
DNA151124 Spike Percents by Reverse Primer



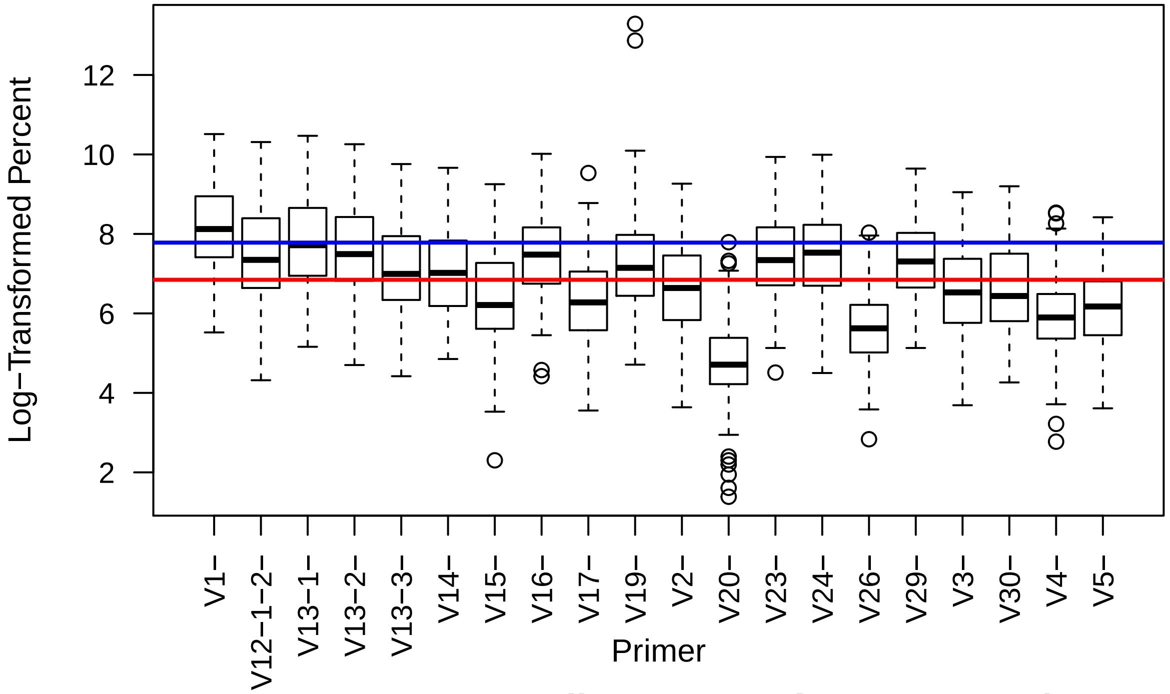
DNA151124 Blood Spike Percents by Forward Primer



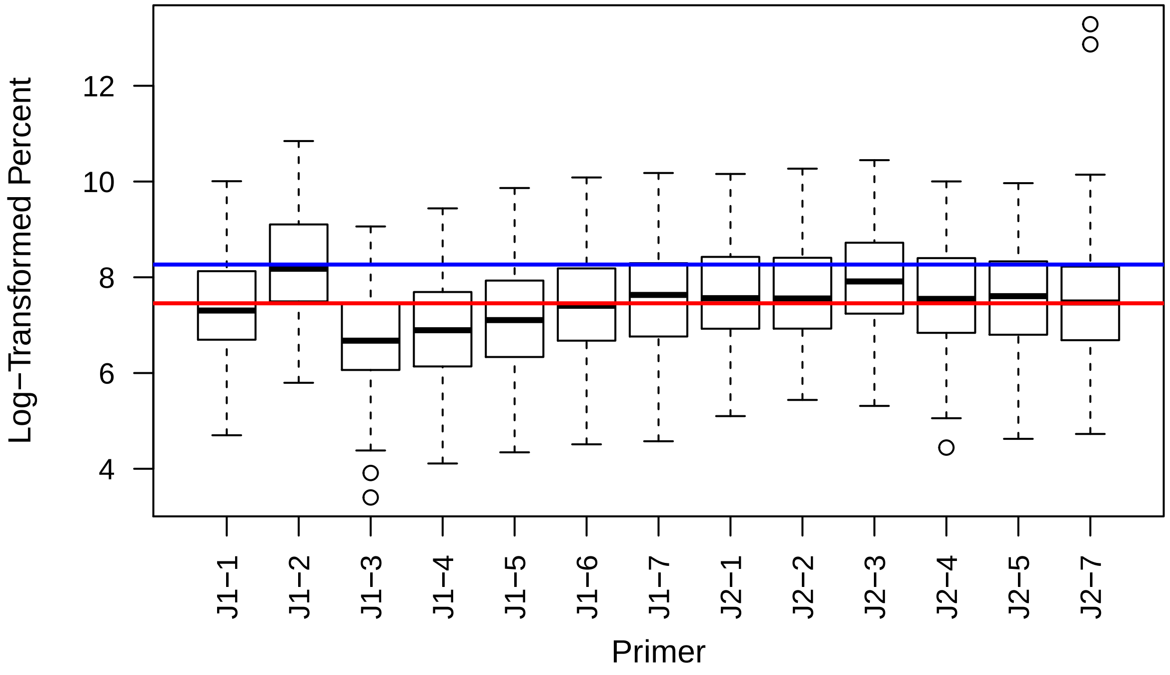
DNA151124 Blood Spike Percents by Reverse Primer

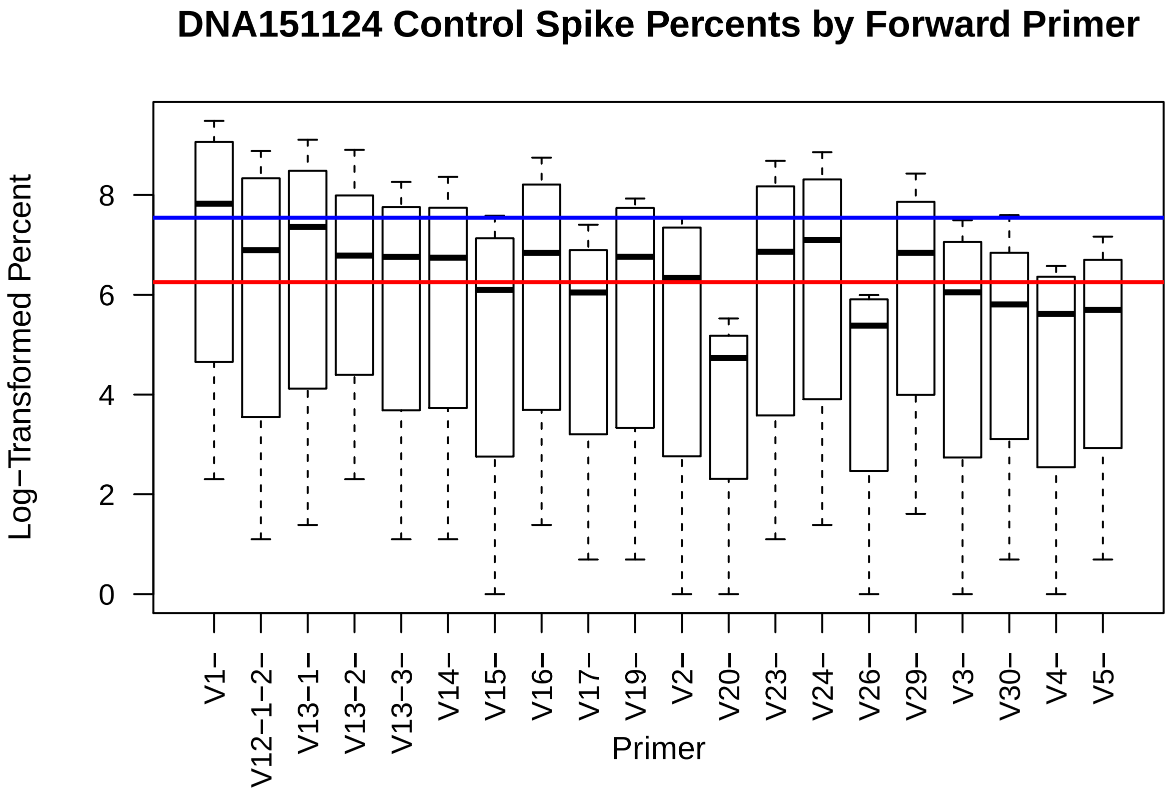


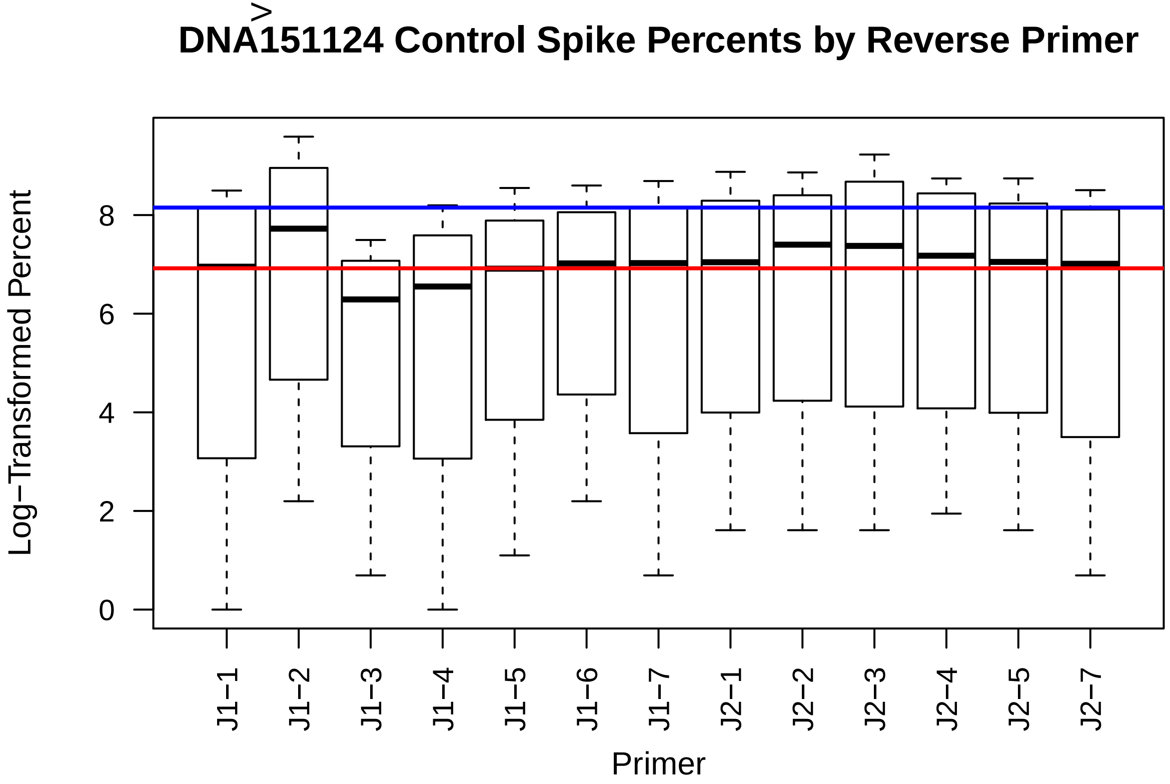
DNA151124 FFPE Spike Percents by Forward Primer



DNA151124 FFPE Spike Percents by Reverse Primer







We can zoom in even further and look at each individual primer combination, instead of grouping by V and J. The plot is too large to fit in this pdf, however, and must be exported.

```
### Melt data frames appropriately
# Full dataset
```

```r
qc.data.25.melt <- melt(qc.data.25.xform, id.vars = c("V", "J"), na.rm = F)
qc.data.25.melt$primer <- paste(qc.data.25.melt$V, qc.data.25.melt$J, sep = '')
qc.data.25.melt$log <- log(qc.data.25.melt$value +1)
# Blood
qc.blood.25.melt <- melt(qc.blood.25.xform, id.vars = c("V", "J"), na.rm = F)
# FFPE
qc.ffpe.25.melt <- melt(qc.ffpe.25.xform, id.vars = c("V", "J"), na.rm = F)
# Control
qc.control.25.melt <- melt(qc.control.25.xform, id.vars = c("V", "J"), na.rm = F)

# Plot
bp.data.25.melt <- ggplot(qc.data.25.melt, aes(x = primer, y = log, group = primer)) +
  geom_boxplot(aes(fill=primer), width = 30, show.legend = F)
pdf(file="~/Desktop/facet_test.pdf", width = 50, height = 60)
bp.data.25.melt + facet_grid(V ~ J) + theme(strip.text.x = element_text(size = 20),
                                            strip.text.y = element_text(size = 20),
                                            axis.text = element_text(size = 20))
dev.off()
```

```
## pdf
##   2
```