testing

Wes Horton

July 14, 2016

Overview

When testing primer independence using samples 1-20 from DNA160609LC, Burcu noticed a four-fold range in spike counts between samples. For example, the spike corresponding to V1/J1-1 in sample 1 could have a count of 10, whereas the same spike has a count of 40 in sample 2. We usually give the sequencing core unequal amounts of DNA in our samples, but these 20 samples have the same initial concentration of spikes. We would expect them to then have the same final concentration as well, assuming that they amplified similarly.

We can compare the 9-bp spike count totals with the PEAR fastq read counts and determine if the distributions are similar.

Set-up

```
### 9-bp counts
spike.count.dir <- "/Volumes/DNA160609LC/spike_counts/9bp/counts/"
spike.count.files <- list.files(spike.count.dir)
spike.count.files <- spike.count.files[order(as.numeric(gsub(".*_S|\\.assemb.*", '', spike.count.files))
spike.count.files <- spike.count.files[1:20]

### PEAR files
### Read in and sort pear files
#pear.dir <- "/Volumes/DNA160609LC/peared_fastqs/counting/"
#pear.files <- list.files(pear.dir)
#pear.files <- pear.files[order(as.numeric(gsub(".*_S|\\.assem.*", '', pear.files)))]
```

Compare distributions

```
# Spike counts of 9-bp counts
spike.counts <- NULL
for (i in 1:length(spike.count.files)){
  curr.spike.count <- read.table(paste(spike.count.dir, spike.count.files[i], sep = ''),
                                 header = T, stringsAsFactors = F, sep = ',')
  curr.count <- curr.spike.count$spike.count[1]
  spike.counts <- c(spike.counts, curr.count)
} # for

# Read counts of raw files
# pear.counts <- NULL
# for (i in 1:length(pear.files)){
#   curr.pear.count <- as.numeric(system(paste("grep -c '^@' ", pear.dir, pear.files[i], sep = ''), int
#   pear.counts <- c(pear.counts, curr.pear.count)
# } # for
# Above takes too long, just did in bash and imported:
pear.counts <- c(686374, 1209673,  865117, 1095477, 1445022, 1100437, 1433990, 1864230,
                 1328501, 1078247, 1916805, 1011097,  768413,  508421,  925440,  716742, 572435, 914924
```

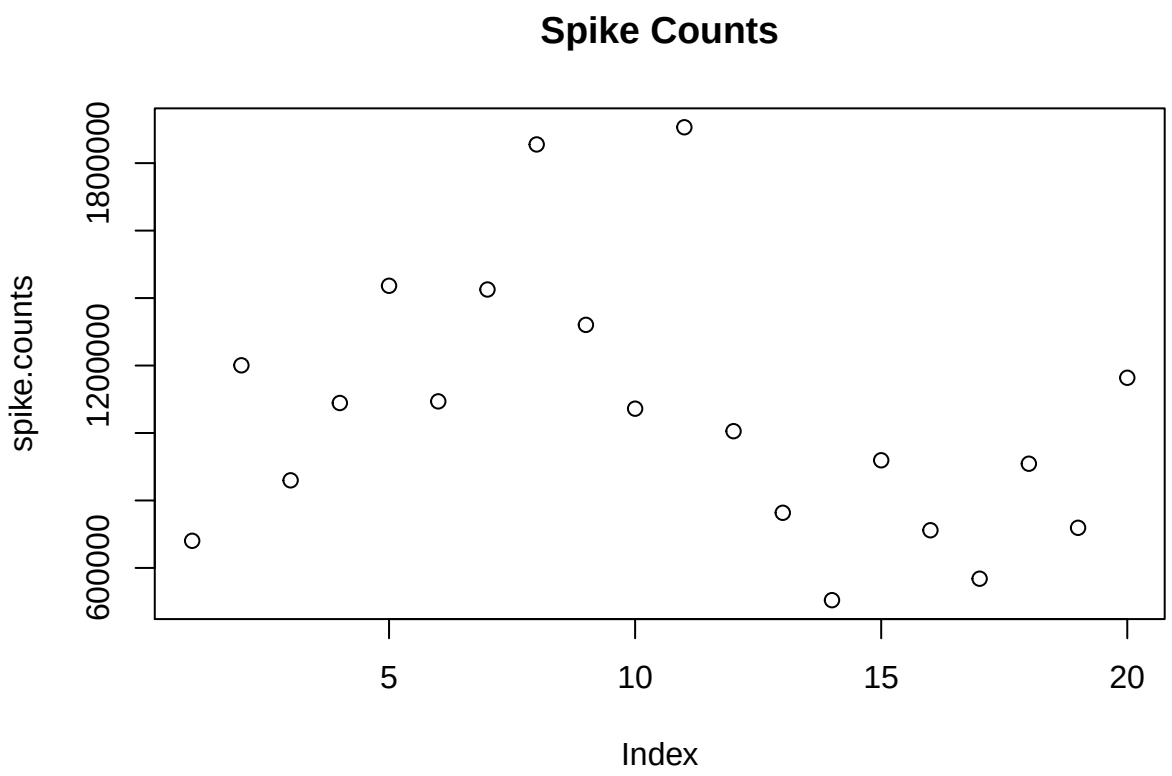
```
                723413, 1169669)

spike.counts
```

```
##  [1]  680413 1200647  859639 1088957 1436461 1093733 1425387 1855464
##  [9] 1320599 1072110 1906327 1005399  763504  504436  919145  711713
## [17]  568063  908939  718989 1163698
```

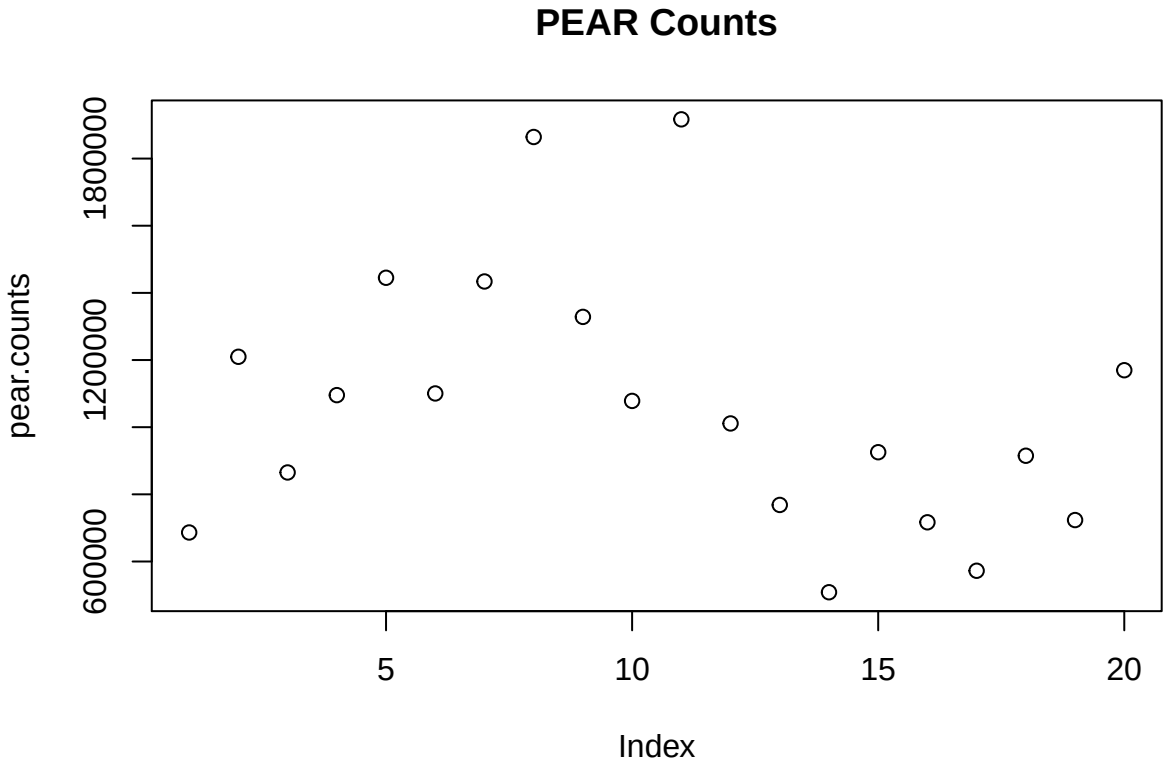
```
pear.counts
```

```
##  [1]  686374 1209673  865117 1095477 1445022 1100437 1433990 1864230
##  [9] 1328501 1078247 1916805 1011097  768413  508421  925440  716742
## [17]  572435  914924  723413 1169669
```

```
plot(spike.counts, main = "Spike Counts")
```




```
plot(pear.counts, main = "PEAR Counts")
```



We can see that the distributions are almost identical. These results suggest that the variation came from any of the following (or a combination of them):

1. Sample preparation - variable amounts of spikes were pipetted into the samples
2. PCR amplification - samples have same starting material, but individual reactions amplified differently in the thermocycler
3. Sequencing - Samples seeded unevenly on the sequencing machine

Next Steps

Do we have quantitative data of sample concentration prior to PCR (nanodrop or cubit?), if so, we should check that distribution. We can also ask Bob if he saw a