DNA160609_spike_only_contamination

Wes Horton

July 14, 2016

Overview

In the control batch DNA160609LC, samples 1-20 contain only spike-ins and primers (no DNA). During our pipeline, we run a spike removal tool that searches for a 9-bp barcode within fastq reads and removes those reads. After running this program on the spike-only samples, we still have reads in these samples.

We need to find out what proportion of total reads are not spikes, as well as what the source of these reads are. One likely cause is from the p14 DNA that contaminated the batch. They could also be spikes that were not removed in the spike removal step.

Set up

We need the PEAR'ed fastq files for samples 1-20. These will act as the baseline counts for each sample. The despiked fastq files produced by the spike removal tool will be our comparison counts. Lastly, we'll need the exported alignment files in order to determine p14 contamination.

Analysis

Using these files, we'll determine

1. The proportion of reads that aren't spikes
2. The proportion of non-spiked reads that are from p14
3. The proportion of reads that aren't spikes, but also are not p14 (unaccounted reads)

Only a small percentage of the total reads are not the spike-ins (less than 1%). Of these non-spike reads, a small amount are accounted for by the p14 contamination, with the rest of unkown origin.

```
##    Prop.not.spikes Prop.of.not.spike.p14 Prop.not.spikes.or.p14
## 1             0.87                  3.77                    0.84
## 2             0.75                  3.47                    0.72
## 3             0.63                  0.42                    0.63
## 4             0.60                  0.48                    0.59
## 5             0.59                  0.41                    0.59
## 6             0.61                  0.76                    0.60
## 7             0.60                  1.09                    0.59
## 8             0.47                  0.75                    0.47
## 9             0.59                  2.10                    0.58
## 10            0.57                  1.94                    0.56
## 11            0.55                  1.65                    0.54
## 12            0.56                  1.14                    0.56
## 13            0.64                  1.34                    0.63
## 14            0.78                  1.05                    0.78
## 15            0.68                  0.95                    0.67
## 16            0.70                  0.46                    0.70
## 17            0.76                    NA                      NA
```

```
## 18            0.65                  0.43                  0.65
## 19            0.61                  0.81                  0.61
## 20            0.51                  1.11                  0.50
```

```
## Proportion of Total Reads that aren't spikes (currently unaccounted for)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.470   0.585   0.610   0.636   0.685   0.870
```

```
## Proportion of Unnacounted Reads that are likely p14 contaminants
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.410   0.615   1.050   1.270   1.495   3.770
```

```
## Proportion of Total Unaccount Reads (Not spikes and not p14 contamination)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4700  0.5700  0.6000  0.6216  0.6600  0.8400
```

Removing p14 contamination doesn't significantly change the proportion of contaminated reads. Are there any primer combinations that stand out as being overrepresented?

```
##    sample total.count Best V hit Best J hit   N proportion
## 1:      1        5531      V13-3       J2-4 246   4.447659
## 2:      1        5531        V20       J2-1 231   4.176460
## 3:      1        5531      V13-1       J1-6  93   1.681432
## 4:      1        5531        V24       J1-6  84   1.518713
## 5:      1        5531      V13-2       J2-1  83   1.500633
## 6:      2        8311      V13-3       J2-4 321   3.862351
```

None of the primer combinations have particularly high proportions. One thing of note, however, is that V13-1, V13-2, and V13-3 seem to appear often. We can group by V's instead of V/J and see if any particular V's are messing things up.

```
##    sample total.count Best V Hit Best.V.Count best.proportion Worst V Hit
## 1:      1        5531      V13-3          591       10.685229         V27
## 2:      1        5531      V13-1          533        9.636594         V22
## 3:      1        5531        V24          482        8.714518       V12-2
## 4:      1        5531      V13-2          438        7.919002         V26
## 5:      1        5531         V1          375        6.779967          V4
## 6:      2        8311      V13-1          871       10.480087         V25
##    Worst.V.Count worst.proportion
## 1:             1       0.01807991
## 2:             4       0.07231965
## 3:            19       0.34351835
## 4:            65       1.17519436
## 5:           123       2.22382933
## 6:             1       0.01203225
```

```
## Top 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   41.70   42.95   43.82   44.20   44.98   48.93
```

```
##    Best V Hit  N
## 1:      V13-3 17
## 2:      V13-1 19
## 3:        V24 19
## 4:      V13-2 19
## 5:         V1 19
## 6:      V12-1  2
```

```
## Low 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2373  1.8500  4.0830  4.1330  5.4890 10.1300
```

```
##     Worst V Hit  N
##  1:         V27  7
##  2:         V22  9
##  3:       V12-2 19
##  4:         V26 16
##  5:          V4  7
##  6:         V25  5
##  7:       V12-3  6
##  8:         V18  4
##  9:         V11  3
## 10:          V5  8
## 11:         V19  2
## 12:         V21  3
## 13:         V20  1
## 14:          V8  2
## 15:          V3  1
## 16:         V15  1
## 17:         V28  1
```

In the alignment files, the same 5 V's are almost always the top 5, and they generally make up about 43.8204509 percent of all of the reads. The V's with the lowest counts are variable, with 17 of the 20 V's in this list at least once. Of note are V12-2 and V26, which are in the low 5 in a majority of samples. Is this the same in the spike count files?

```
##    sample total.count Best V Hit best.v.count top.proportion Worst V Hit
## 1:      1      632155      V13-1        52895       8.367410         V15
## 2:      1      632155         V1        52794       8.351433          V5
## 3:      1      632155        V29        44537       7.045266          V4
## 4:      1      632155        V24        44069       6.971233         V20
## 5:      1      632155        V23        40251       6.367268         V26
## 6:      2     1134018         V1        95528       8.423852         V15
##    worst.v.count worst.proportion
## 1:         22684         3.588360
## 2:         18086         2.861007
## 3:         17553         2.776692
## 4:         15608         2.469015
## 5:         14025         2.218601
## 6:         41492         3.658848
```

```
##    sample total.count Best J Hit best.j.count top.proportion Worst J Hit
## 1:      1      632155       J1-2        91116      14.413554        J2-5
## 2:      1      632155       J1-7        60081       9.504156        J1-5
## 3:      1      632155       J2-3        58796       9.300883        J2-7
## 4:      1      632155       J1-1        53881       8.523384        J1-4
## 5:      1      632155       J2-1        52751       8.344631        J1-3
## 6:      2     1134018       J1-2       162990      14.372788        J2-5
##    worst.j.count worst.proportion
## 1:         44004         6.960951
## 2:         36969         5.848091
## 3:         35200         5.568255
## 4:         32476         5.137348
## 5:         21874         3.460227
## 6:         78717         6.941424
```

```
## Top 5
```

```
##    Best V Hit  N
## 1:      V13-1 19
## 2:         V1 19
## 3:        V29 19
## 4:        V24 19
## 5:        V23 16
## 6:      V13-2  3
```

```
## Worst 5
```

```
##    Worst V Hit  N
## 1:         V15 19
## 2:          V5 19
## 3:          V4 19
## 4:         V20 19
## 5:         V26 19
```

```
## Top 5
```

```
##    Best J Hit  N
## 1:       J1-2 19
## 2:       J1-7 19
## 3:       J2-3 19
## 4:       J1-1 19
## 5:       J2-1 19
```

```
## Worst 5
```

```
##    Worst J Hit  N
## 1:        J2-5 14
## 2:        J1-5 19
## 3:        J2-7 19
## 4:        J1-4 19
## 5:        J1-3 19
## 6:        J2-4  5
```

The unique V's that are in the top 5 in at least one sample are not the same between the left-over reads and the spiked reads, although 3 of them are the same. The low 5 are the same in all of the spiked samples. In fact, V13-1, V1, and V24 are in the top 5 in all samples.

Currently, we identify spikes for removal using the first 9-bp sequence only. We could potentially use both the first and the second 9-bp sequences to remove spikes. After implementing the new spike finding, we can compare the identified spikes between the two techniques.

```
##  [1] 3038 4248 2103 2582 3325 2654 3558 3307 3484 2487 4168 2400 2154 1906
## [15] 2552 1930 1894 2292 1709 2275
```

```
##  [1] 50.96 47.06 38.39 39.60 38.84 39.59 41.36 37.73 44.09 40.52 39.78
## [12] 42.12 43.88 47.83 40.54 38.38 38.30 38.63 38.10
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1774  0.2332  0.2431  0.2631  0.2725  0.4426
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4700  0.5700  0.6000  0.6216  0.6600  0.8400
```

```
##    Sample Orig.Prop.Not.Spikes Orig.Unaccounted Prop.Not.Spikes
## 1       1                0.87%            0.84%         0.4426%
## 2       2                0.75%            0.72%         0.3512%
## 3       3                0.63%            0.63%         0.2431%
## 4       4                 0.6%            0.59%         0.2357%
## 5       5                0.59%            0.59%         0.2301%
## 6       6                0.61%             0.6%         0.2412%
## 7       7                 0.6%            0.59%         0.2481%
## 8       8                0.47%            0.47%         0.1774%
## 9       9                0.59%            0.58%         0.2623%
## 10     10                0.57%            0.56%         0.2307%
## 11     11                0.55%            0.54%         0.2174%
## 12     12                0.56%            0.56%         0.2374%
## 13     13                0.64%            0.63%         0.2803%
## 14     14                0.78%            0.78%         0.3749%
## 15     15                0.68%            0.67%         0.2758%
## 16     16                 0.7%             0.7%         0.2693%
## 17      1                0.65%            0.65%         0.2505%
## 18      2                0.61%            0.61%         0.2362%
## 19      3                0.51%             0.5%         0.1945%
##    Prop.Unaccounted
## 1           0.4098%
## 2           0.3253%
## 3           0.2404%
## 4           0.2329%
## 5           0.2277%
## 6           0.2365%
## 7           0.2416%
## 8           0.1739%
## 9           0.2498%
## 10          0.2196%
## 11          0.2084%
## 12          0.2309%
```

```
## 13          0.2717%
## 14          0.3666%
## 15          0.2693%
## 16          0.2661%
## 17          0.2477%
## 18          0.2313%
## 19          0.1889%
```

Now a very small amount of the data are not spikes or p14 (i.e. unaccounted for), but we still don't know where it comes from. I suggest that we BLAST these reads and try and see if they match to anything known.

```
##    Fasta.reads Blast.reads
## 1         3034        2347
## 2         4248        3334
## 3         2101        1656
## 4         2579        2086
## 5         3325        2599
## 6         2653        2079
## 7         3558        2789
## 8         3306        2555
## 9         3482        2681
## 10        2487        1979
## 11        4165        3277
## 12        2399        1844
## 13        2152        1441
## 14        1905        1512
## 15        2551        2085
## 16        1928        1517
## 17        1893        1494
## 18        2291        1800
## 19        1708        1358
## 20        2274        1819
```