Deeper Sequencing Analysis

Madison Hobbs

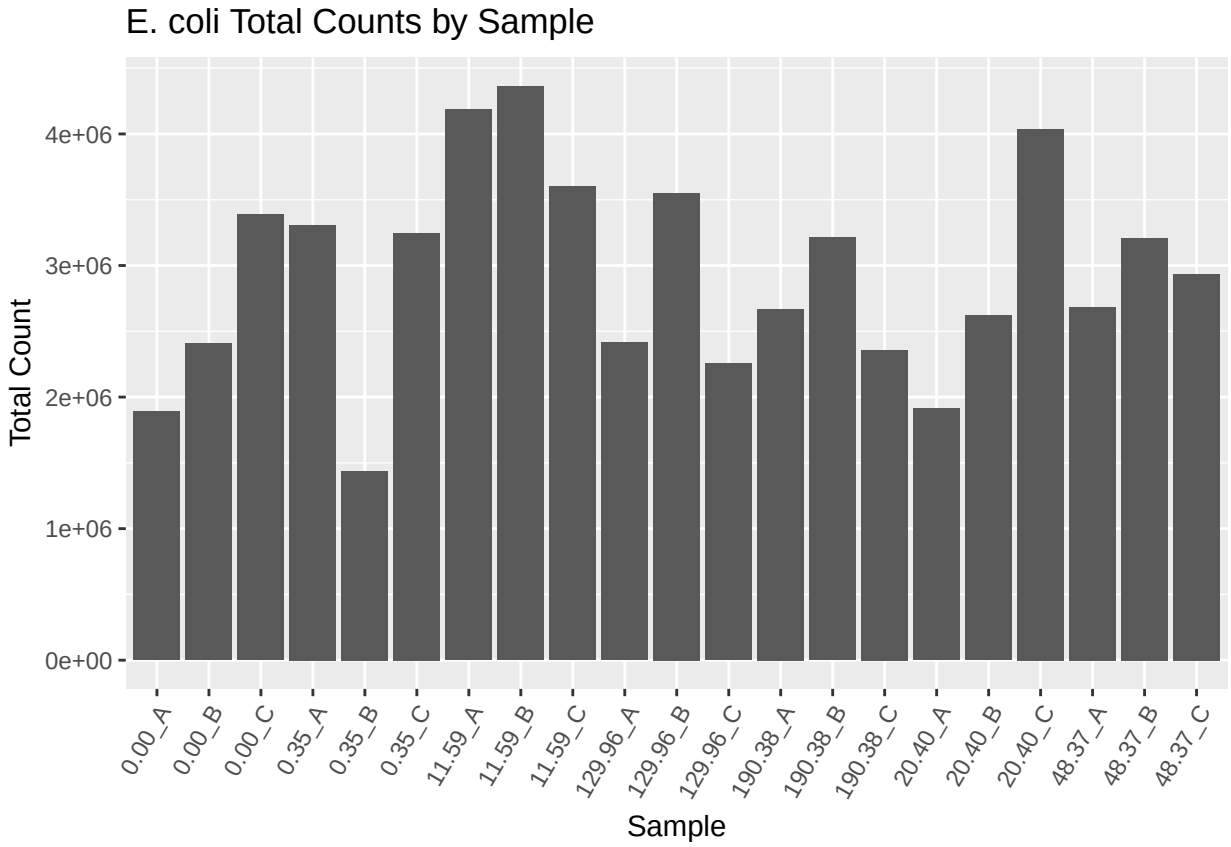
1/10/2018

Note: for ease, "deep" data refers to the data just sequenced more deeply while "shallow" data refers to the data we worked on this summer.

The following document compares some diagnostics between the shallow data and deep data, then visualizes the gene expression patterns of both datasets in a clustering setting.

Total Counts and Size Factors

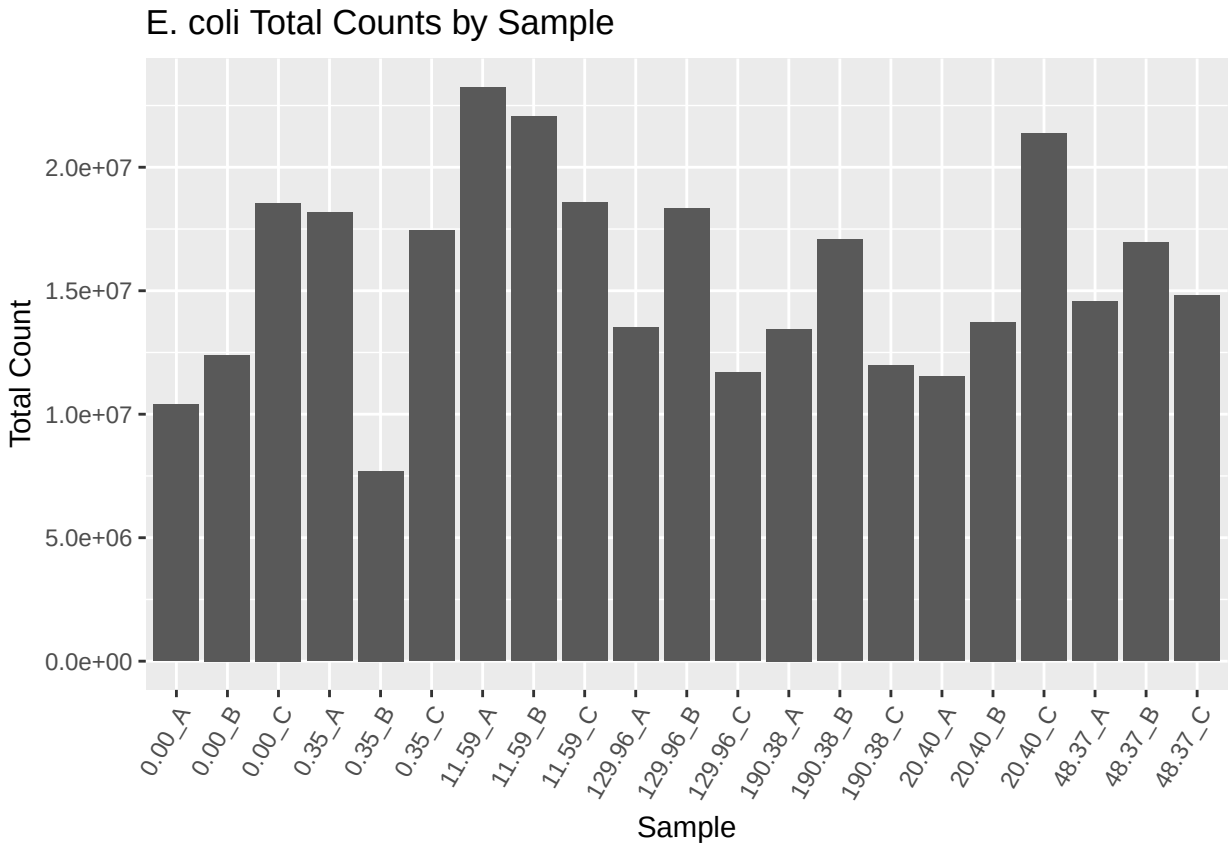
Shallow Data




```
## # A tibble: 21 x 6
##    condition sample median    Q3 sizeFactor totalCount
##        <dbl>  <chr>  <dbl> <dbl>      <dbl>      <dbl>
##  1      0.00      A      4    20  0.8873935    1889658
##  2      0.00      B      1     4  0.1828702    2409296
##  3      0.00      C      3    14  0.6212055    3386987
##  4      0.35      A      7    29  1.3174901    3308303
##  5      0.35      B      1     3  0.1328494    1436978
##  6      0.35      C      6    28  1.2531979    3246960
```

```
##  7     11.59      A      8    40  1.7143554    4187736
##  8     11.59      B      9    40  1.7001676    4362489
##  9     11.59      C      8    35  1.5215741    3600378
## 10     20.40      A      2     8  0.3463918    1915087
## # ... with 11 more rows
```

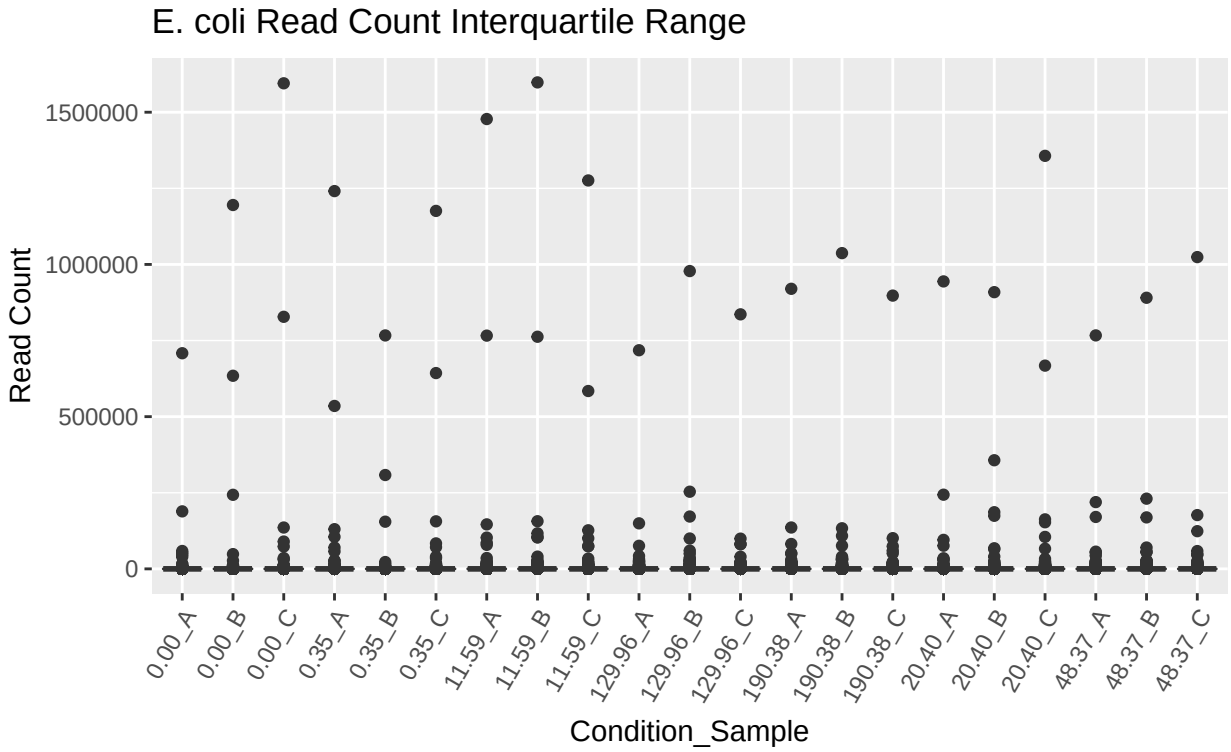
Deep Data

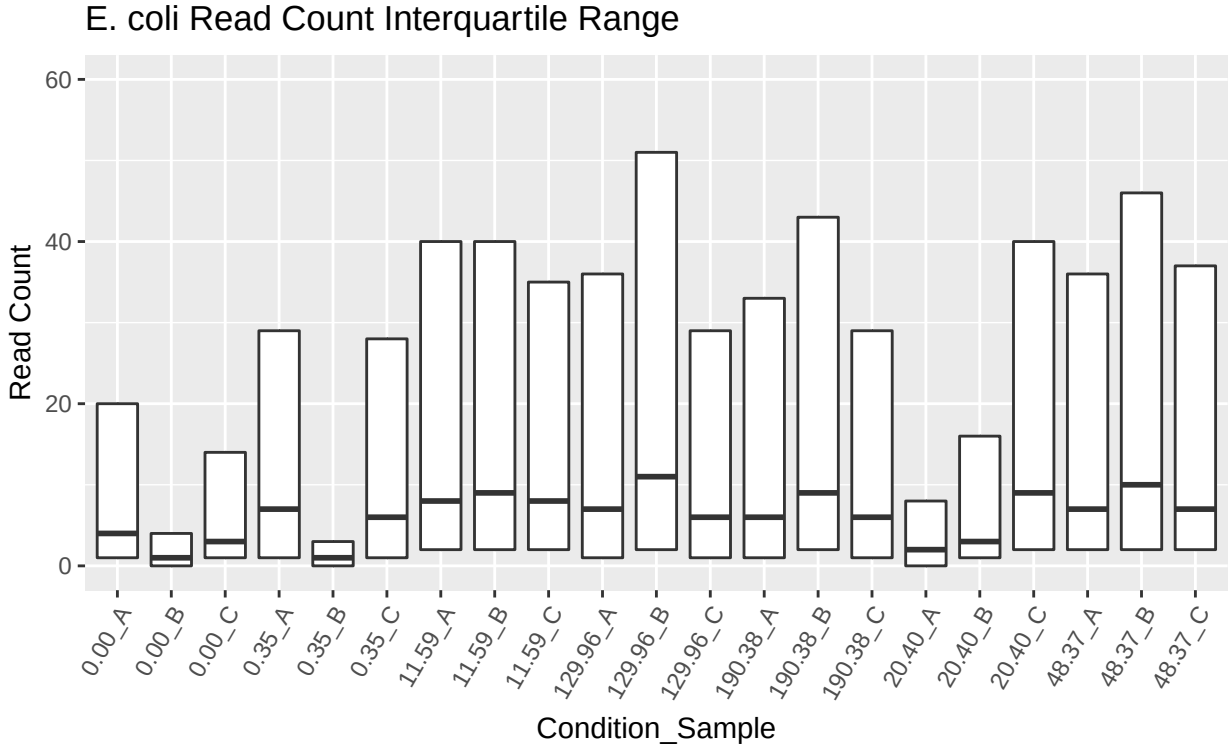



```
## # A tibble: 21 x 6
##    condition sample median    Q3 sizeFactor totalCount
##        <dbl>  <chr>  <dbl> <dbl>      <dbl>      <dbl>
##  1      0.00      A      4    20  0.9034148   10395764
##  2      0.00      B      1     4  0.1760866   12391466
##  3      0.00      C      3    14  0.6412187   18554337
##  4      0.35      A      7    29  1.3468605   18179978
##  5      0.35      B      1     3  0.1234918    7703955
##  6      0.35      C      6    28  1.2823126   17446114
##  7     11.59      A      8    40  1.7398799   23240163
##  8     11.59      B      9    40  1.6514763   22053814
##  9     11.59      C      8    35  1.4826783   18570454
## 10     20.40      A      2     8  0.3626991   11522258
## # ... with 11 more rows
```

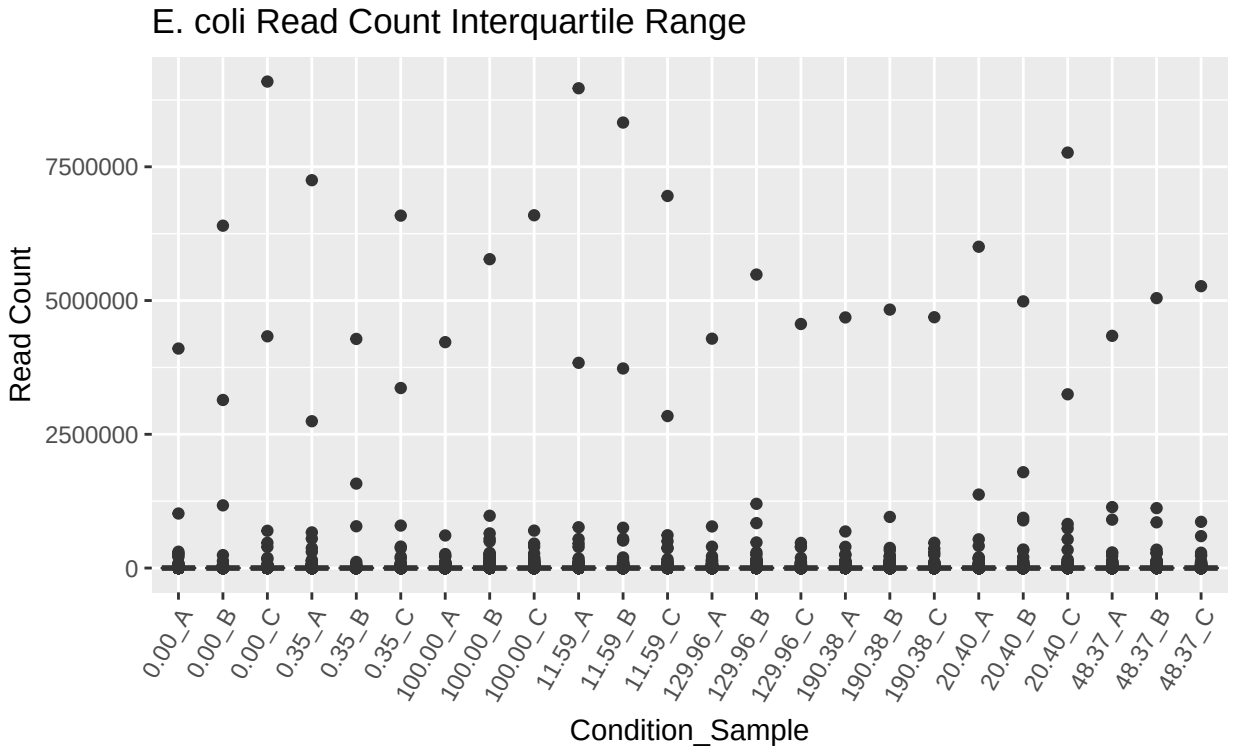
We see, unsurprisingly, that the total count per sample of the deeper sequenced data. is larger than that of the shallower sequenced data. However, the size facotors are similar because the distribution of total counts across samples is similar for both depths of sequencing.

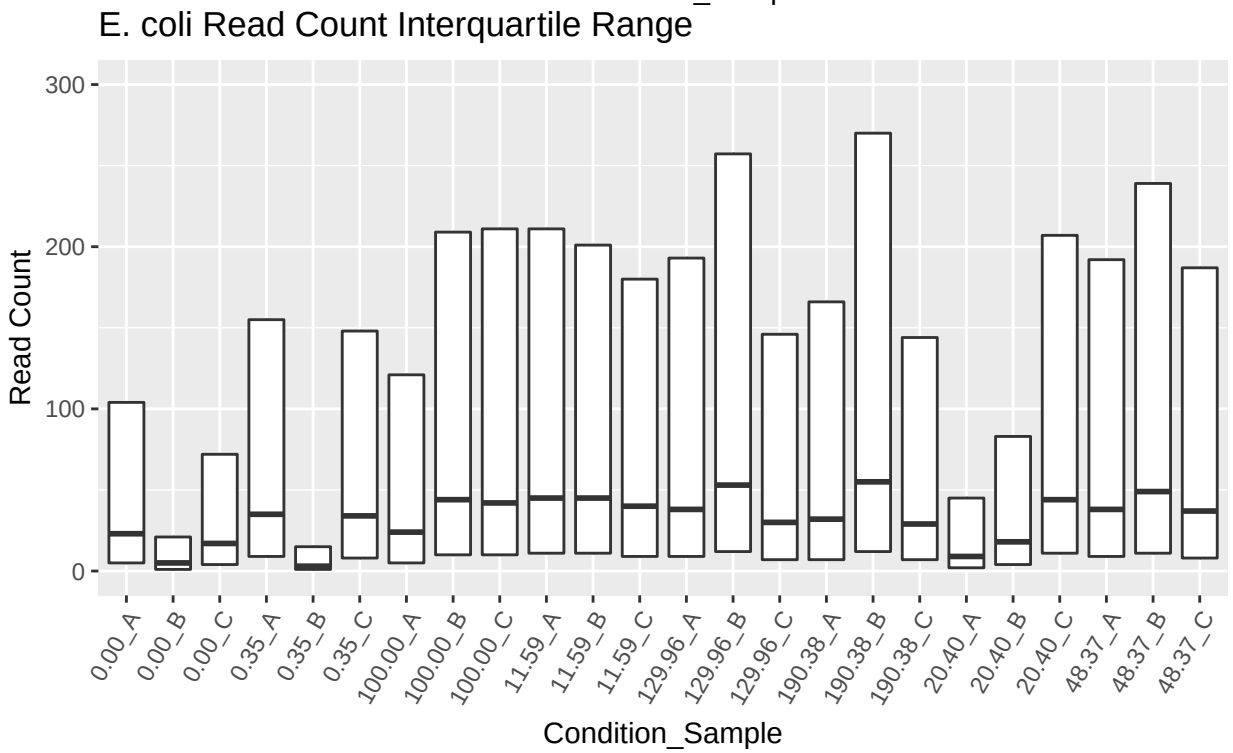
Shallow Data





Deep Data





Note that the y axes on both graphs have different scales, but the distribution of readcounts per sample look similar across the two sequencing depths. the samples with majority lower read count genes in the shallow data still display majority lower read count in the deep data. That is, the samples with the lowest Q3 read count in the shallow data are the same as the samples with lowest read count in the deep data: namely,

0.00_B, 0.35_B, and 20.40_A. The rest of the samples in the deep data follow a similar distribution to those in the shallow data.

This result could suggest that the deep data is confirming what the shallow data saw, but gene by gene plots are needed to see if this is true.

Before, we eliminated genes for which the maximum unnormalized count of any condition was less than 5. In the shallower sequenced data, there were 3,183 such genes. In the deeper data, we have 938 such genes. This makes sense because the deeper sequenced data is more deeply sequenced!

```
## [1] 938.0000000   0.5833333
```

Dominating Genes

Shallow Data

Deep Data

Of the ten most expressed genes from each level of sequencing, eight are the same in both levels of sequencing. These are smpB-intA, rna69, rna106, rmf b0953, rna98, cspE, rna54, and rna71.

```
## [1] 844   3
```

```
## [1] 156   2
```

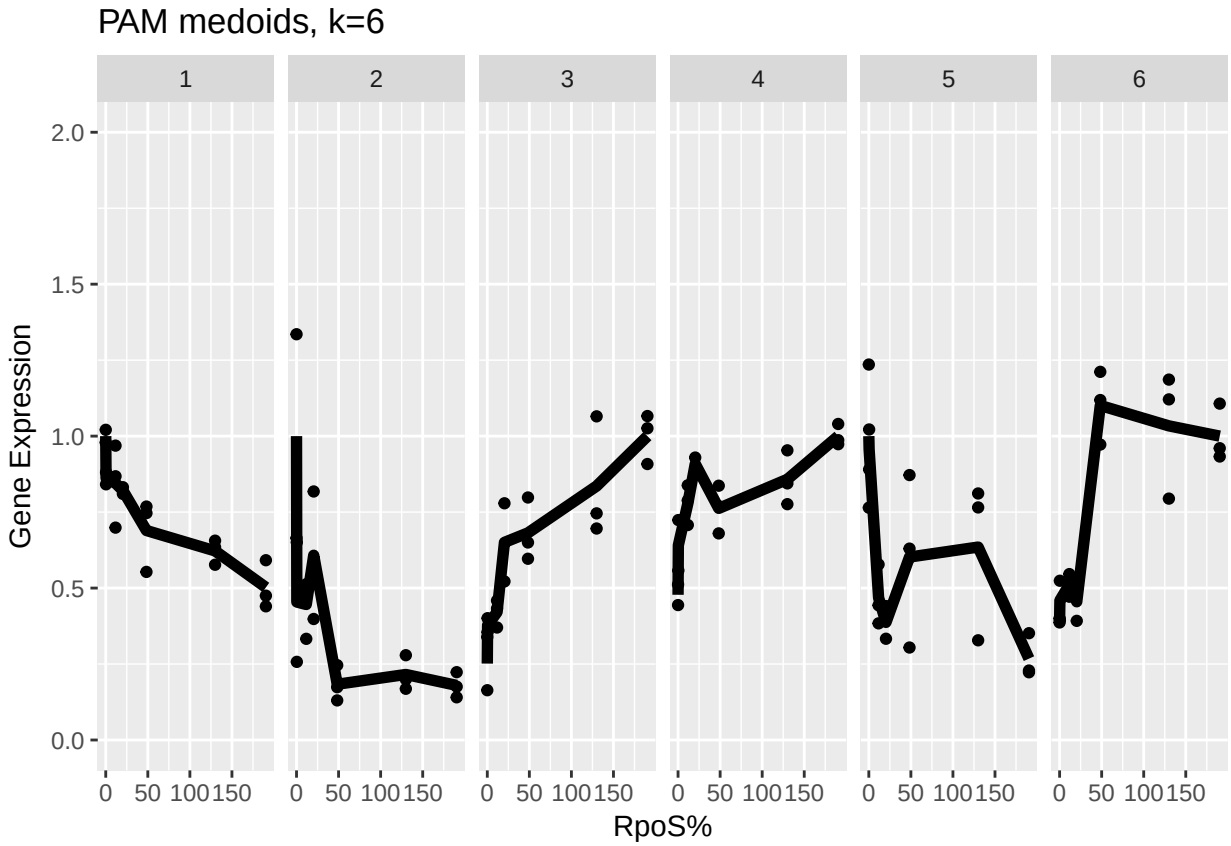
In fact, of the 1000 most expressed genes in each level of sequencing, 844/1000 are the same and 156/1000 are different. This shows that the genes which dominated the shallow data tend to be the same genes dominating the deep data.

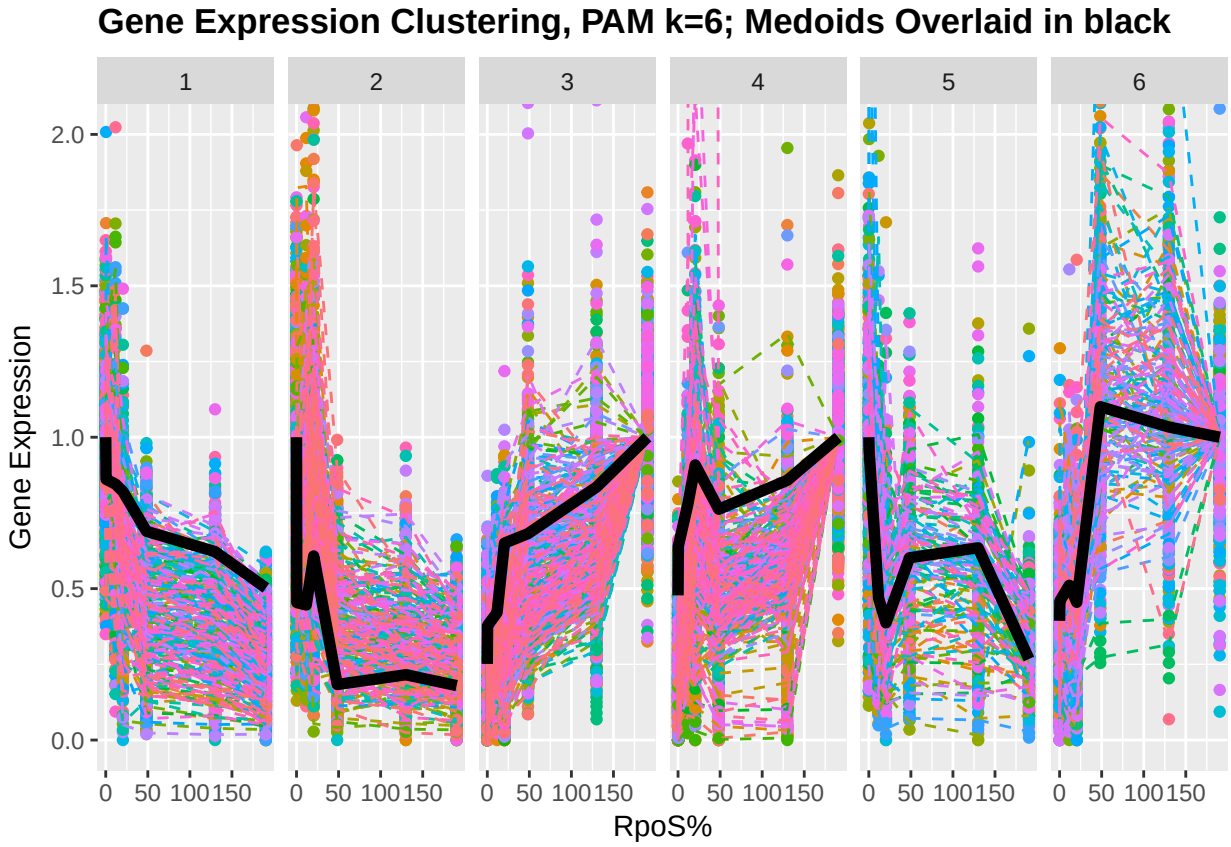
Differential Expression, Gene Expression Shapes, and Clustering

We throw out the samples with much lower median read count (0.00_B, 0.35_B, and 20.40_A) as well as the samples with 100% RpoS (100.00_A, 100.00_B, 100.00_C) because it has a different strain of E. coli than the rest of the samples.

Shallow Data

```
##
## out of 10374 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 1056, 10%
## LFC < 0 (down)   : 793, 7.6%
## outliers [1]     : 1, 0.0096%
## low counts [2]   : 2816, 27%
## (mean count < 5)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

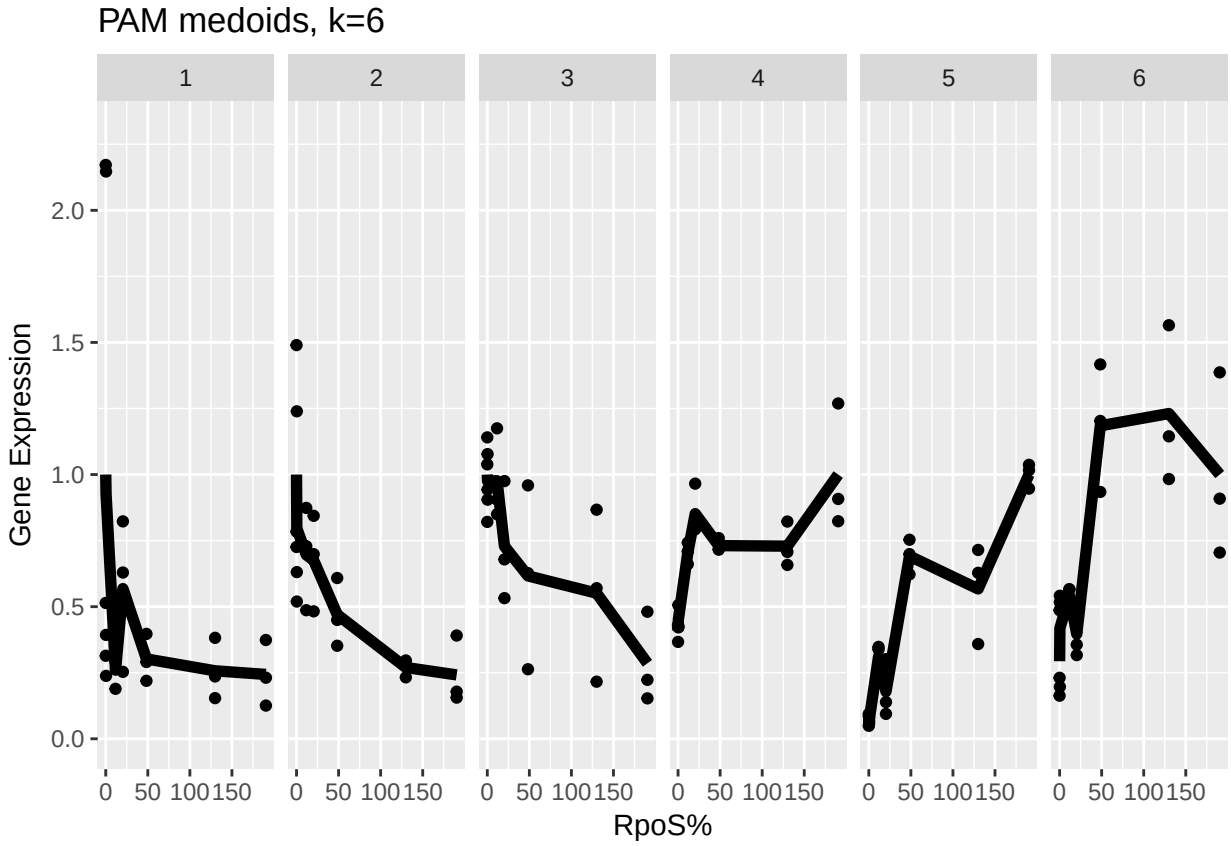

PAM medoids, k=6
1
2
3
4
5
6
Gene Expression
2.0
1.5
1.0
0.5
0.0
0
50
100
150
RpoS%

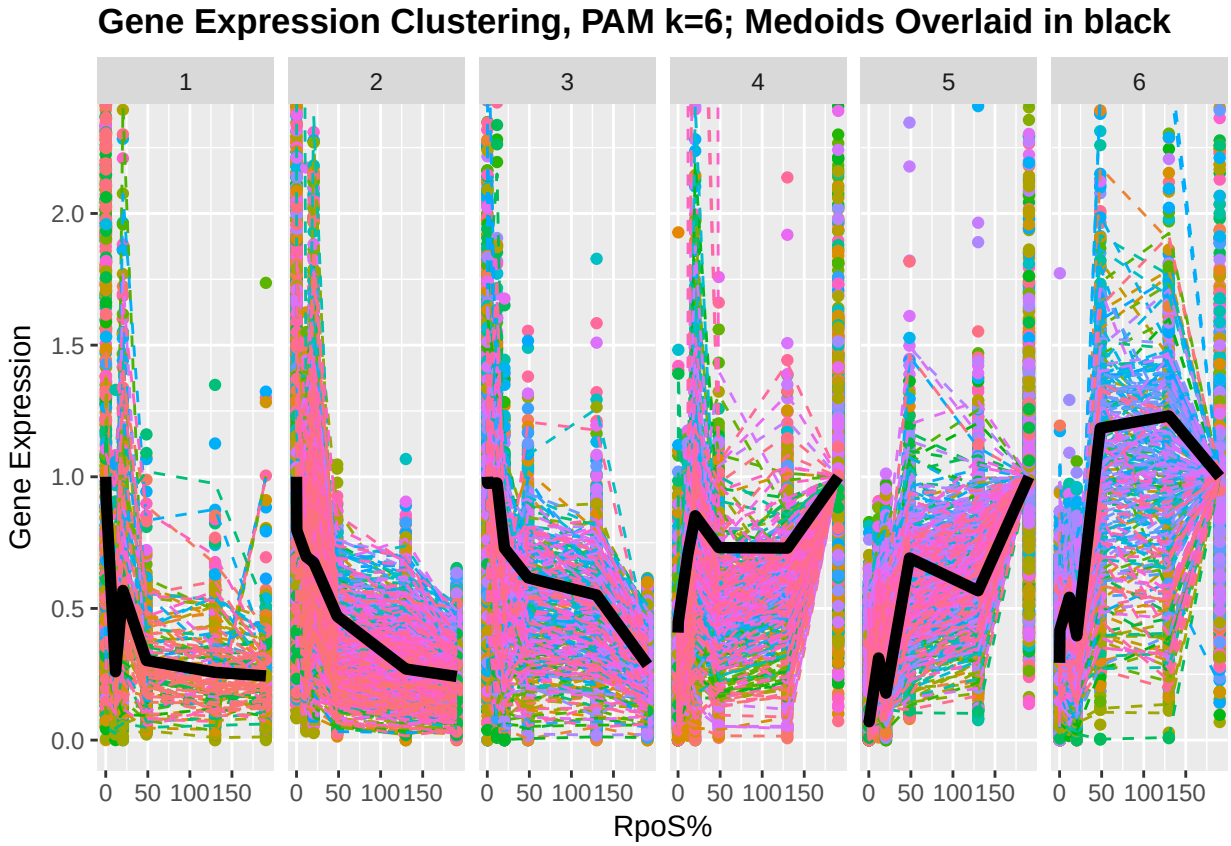

Gene Expression Clustering, PAM k=6; Medoids Overlaid in black
1
2
3
4
5
6
Gene Expression
2.0
1.5
1.0
0.5
0.0
0
50
100
150
RpoS%

Deep Data, all samples except 100% RpoS

.

```
##
## out of 12989 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)      : 1471, 11%
## LFC < 0 (down)    : 1195, 9.2%
## outliers [1]      : 360, 2.8%
## low counts [2]    : 2743, 21%
## (mean count < 8)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```



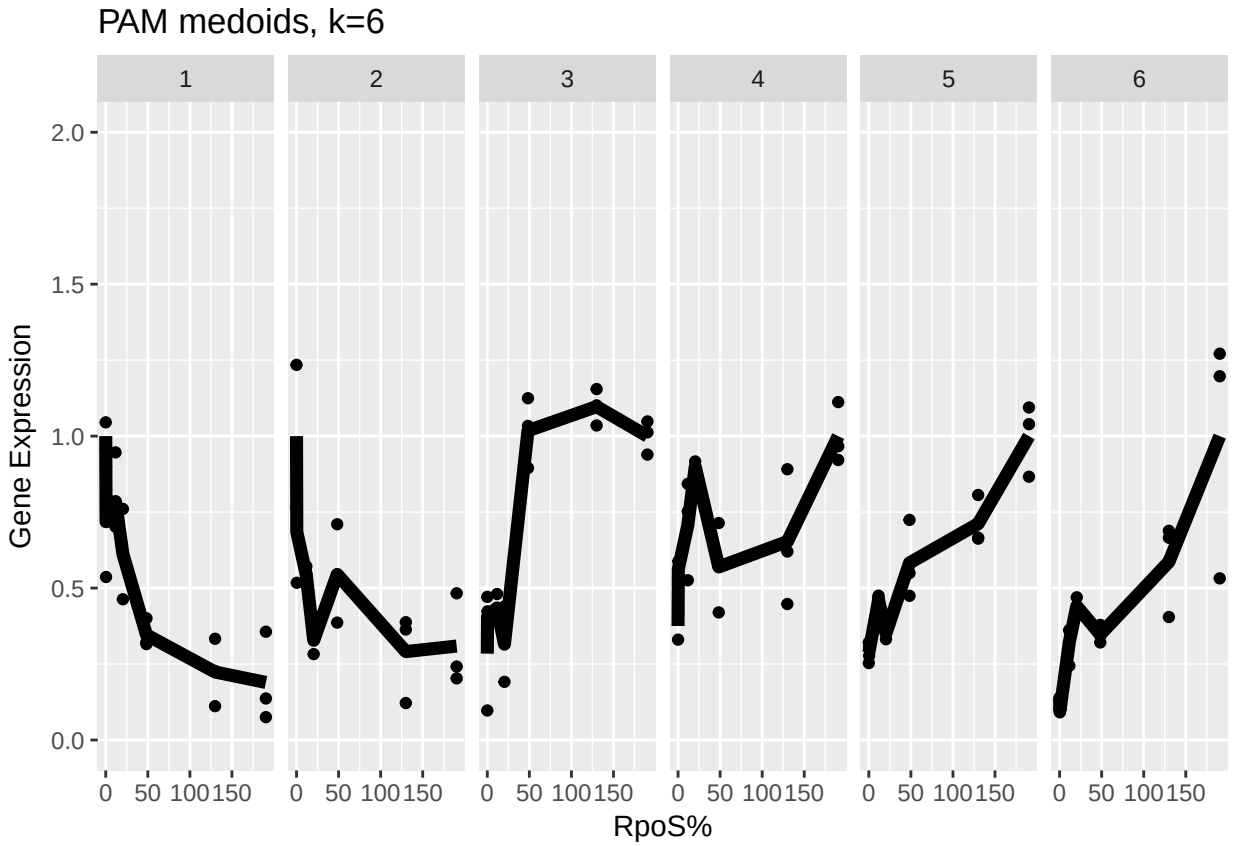


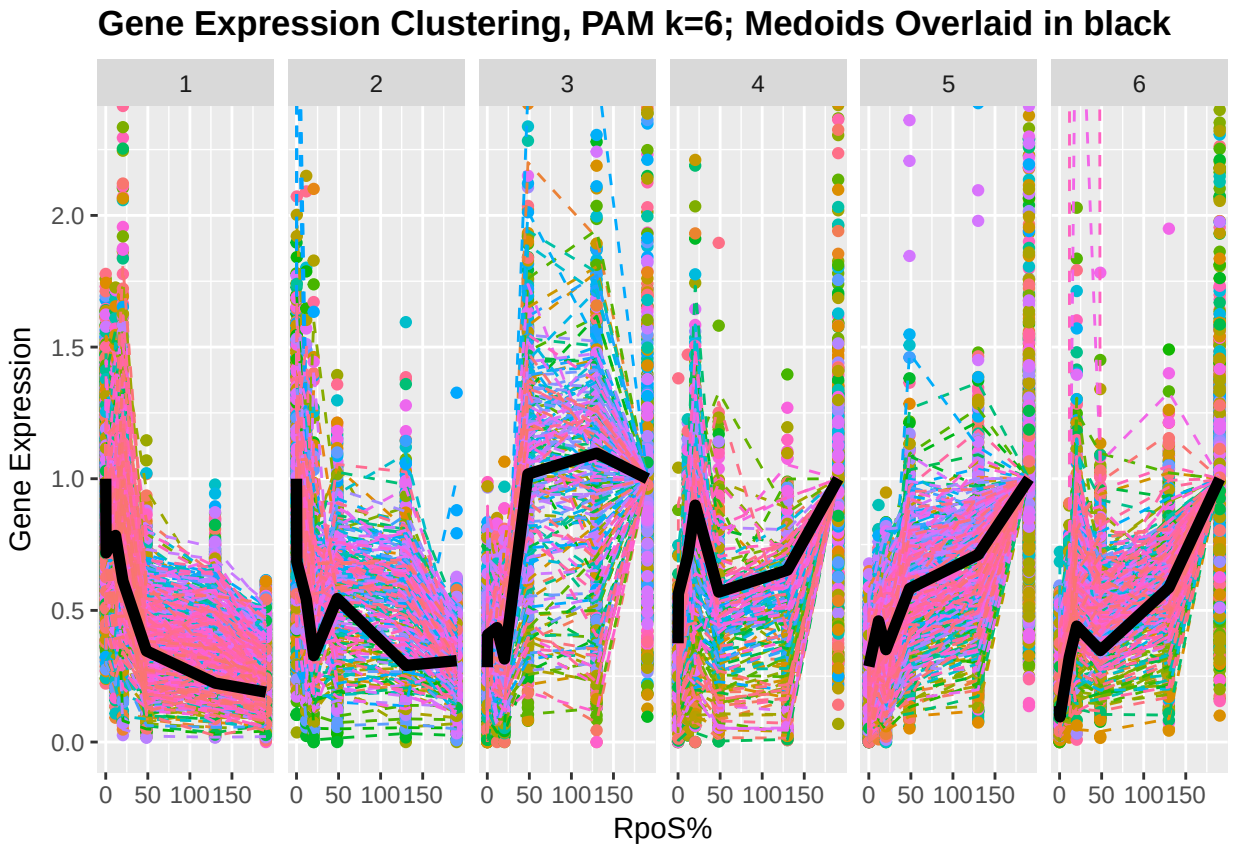
Deep Data, 0.00_B, 0.35_B, 20.40_A and 100% RpoS removed

We remove the samples with 100% RpoS and 0.00_B, 0.35_B, 20.40_A.

.

```
##
## out of 12989 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)      : 1327, 10%
## LFC < 0 (down)    : 992, 7.6%
## outliers [1]      : 219, 1.7%
## low counts [2]    : 2516, 19%
## (mean count < 9)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```


PAM medoids, k=6
1
2
3
4
5
6
Gene Expression
2.0
1.5
1.0
0.5
0.0
0
50
100
150
RpoS%


Gene Expression Clustering, PAM k=6; Medoids Overlaid in black
1
2
3
4
5
6
Gene Expression
2.0
1.5
1.0
0.5
0.0
0
50
100
150
RpoS%

In the shallow and deep data, a similar proportion of genes are found to be differentially expressed across the most extreme levels, 0% and 190.38% RpoS.

In both the shallow and deep data, we notice similar "common shapes" (medoids) and a considerable amount of nonmonoticity in the gene expression across different levels of RpoS.

How Similar are the Deep and Shallow Clusterings?

ARI: Deep (0.00_B, 0.35_B, 20.40_A removed) vs. Shallow:

```
## [1] 0.529
```

The adjusted rand index suggests above that the clusterings for $k = 6$ of the deep data (with 0.00_B, 0.35_B, 20.40_A removed) and shallow datasets are more similar than dissimilar, but certainly don't agree wholeheartedly.

ARI: Deep (0.00_B, 0.35_B, 20.40_A removed) vs. Deep (0.00_B, 0.35_B, 20.40_A included)

```
## [1] 0.603
```

The clusterings for the deep data with and without 0.00_B, 0.35_B, and 20.40_A are more similar than the deep vs. shallow, but certainly don't agree wholeheartedly either.