Questions Regarding 2017 Data

Professor Jo Hardin, Pomona College Mathematics Department; Professor Dan Stoebel, Harvey Mudd College Biology Department; Madison Hobbs, Scripps College

6/27/2017

Table of Contents

0. Our Questions
1. Total Counts for E. coli and B. subtilis
2. Boxplots of Read Counts by Sample
3. Highly Expressed Genes in E. Coli and B. subtilis Data Sets
4. Size Factors
5. Correlations Between Replicates
6. Proportion of Reads From Each Feature in the E. coli data set

0. Our Questions

After some exploring of the data, we have some concerns and would like to seek your expert opinion. Our primary questions for you are the following:

1) Does the sequencing depth seem deep enough, and would higher sequencing depth produce better data?

2) Is the variation of medians and thrid quartiles of the raw counts across each sample a cause for concern? We wonder why the raw read counts for the vast majority of genes are so low, and how we can distinguish actual counts from noise at reads this low.

3) Do the few genes with very large raw read counts suggest a problem?

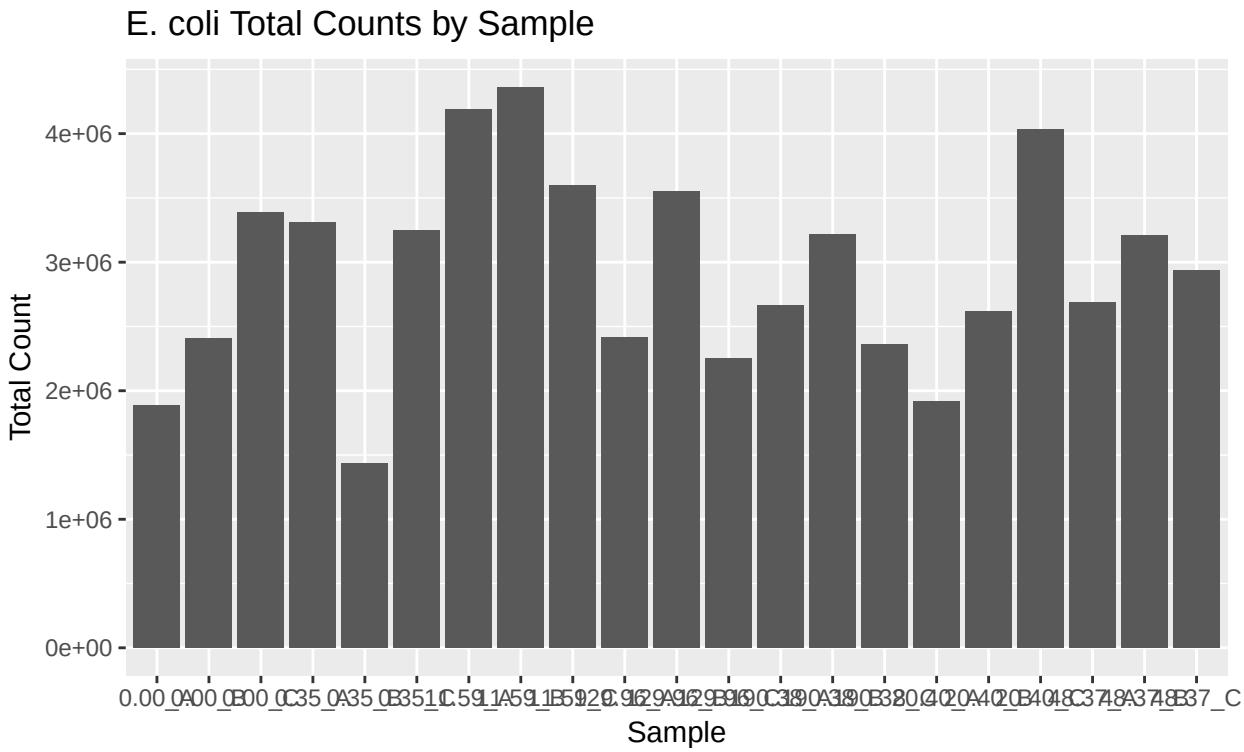
4) Does the spread of size factors suggest a problem, or is this a typical distribution?

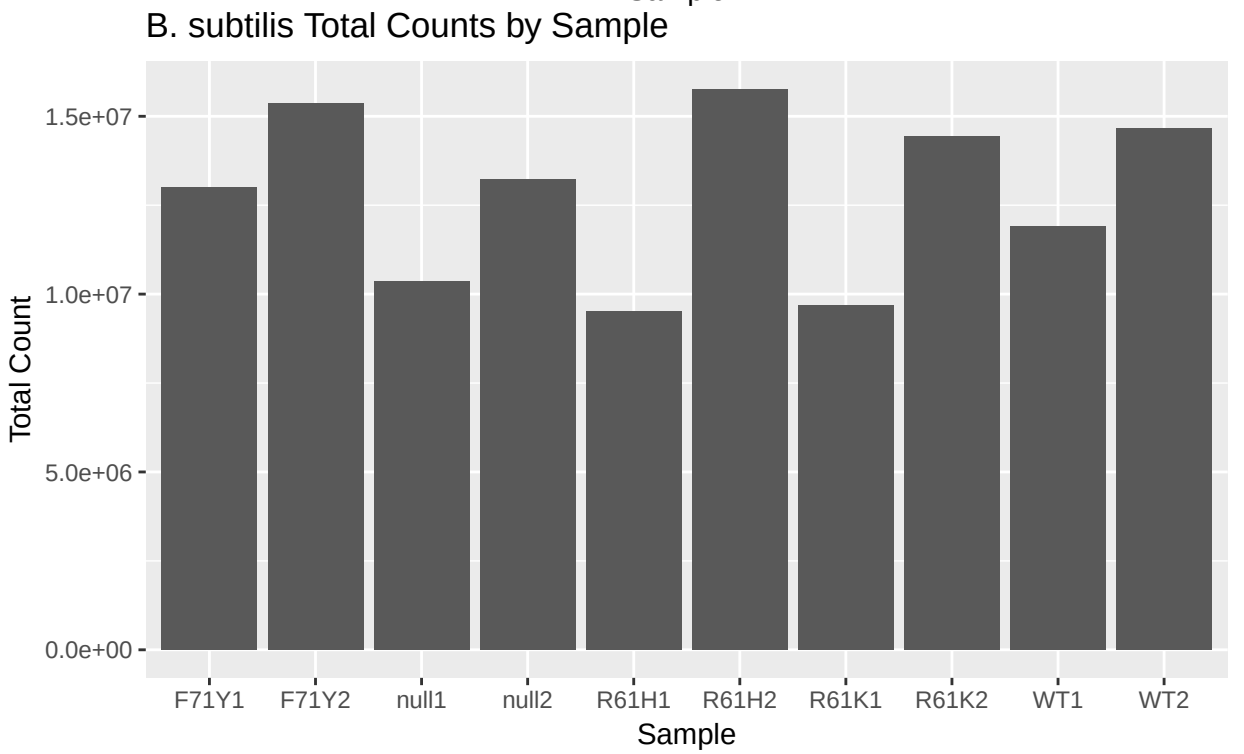
5) Do the correlations between replicates at each condition suggest that the replicates are not true replicates?

6) Is it a problem that the proportion of genes from each feature (CDS, IGR, ncRNA, tRNA, rRNA) varies across the different samples?

By comparing aspects of our data, E. Coli, to the data in the paper on which you collaborated, "Hierarchical expression of genes controlled by the Bacillus subtilis global regulatory protein CodY," we seek to understand if what we see in the E. coli data is expected or reason for concern. Unless otherwise specified, E. Coli data used in this analysis includes all gene features (sens and antisens CDS, IGR, ncRNA, tRNA, rRNA). We would greatly appreciate your perspective and input, and thank you for all of your help, time, and consideration on this project.

1. Total Counts

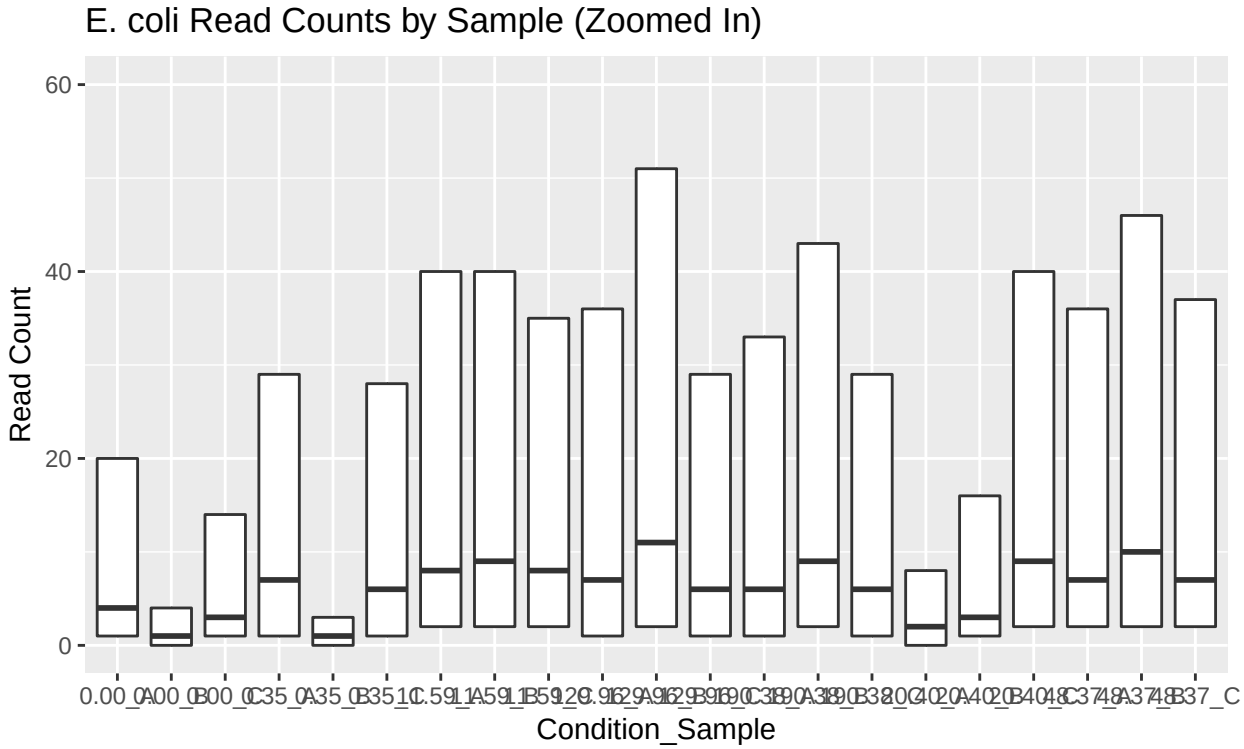




The plots above show the total count (unnormalized) for each sample of the E. coli and B. subtilis data sets respectively. The B. subtilis data set has higher counts across the board, and their range is 6,249,673 and the maximum being around 1.67 times the minimum. The E. coli data has lower and more variation in its total counts by sample, with a range of 2,925,511 and the maximum being around 3.04 times the minimum.

2. Boxplots of Read Counts by Sample

2a. Medians and Third Quartiles of E. Coli and B. subtilis counts

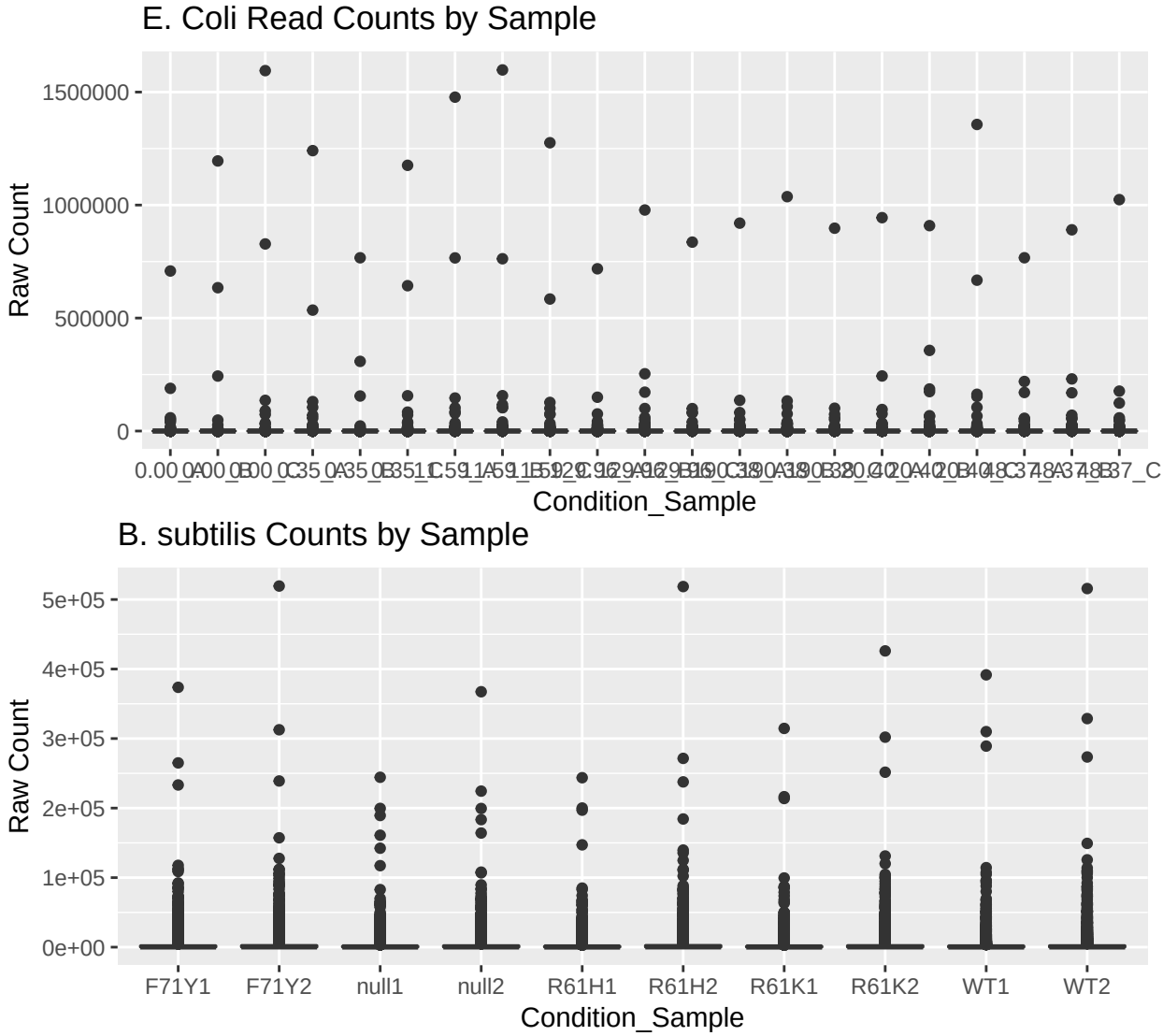


In the above boxplot, the x-axis represents each of the 21 samples, and the y-axis represents the read count (unnormalized) for each sample. This plot is zoomed in on the y-axis so that we can the medians and third quartiles of read counts for each sample. The read counts for the E. coli data display much more variability

across the different samples than we see in the B. subtilis data; the medians and thrid quartiles both vary greatly across different samples (see footnote 1). The B. subtilis data set has median read counts of around 400 for each sample, and third quartiles in the thousands. The vast majority of genes in E. coli data set, on the other hand, have read counts below 40 for each sample, with third quartiles ranging from 3 to 51 and sample median read counts between 1 and 11.

Thus, recalling the total counts from above, in the E. coli data set there exist huge differences between the third quartiles and the maximum of each sample. We find that a few genes dominate the read counts for each sample in the E. coli data.

2b. Full Boxplots of E. Coli and B. subtilis counts



Looking at the boxplots of E. coli and B. subtilis read counts by sample with unrestrained y-axis, we see that they both have a number of genes with read counts higher than sample third quartiles. We notice with E. Coli, however, that the spread between the third quartiles and the highly expressed genes in each sample is much larger.

In the E. coli data, at sample 48.37_C there is one gene at around 200,000 raw counts and the next largest gene is at over 1,000,000 raw counts; at least a five-fold increase. This is not an atypical distance between

consecutive reads in E. coli, nor is it the largest distance. At the same time, the third quartile for 48.37_C is 37, which is a typical, if not relatively high, third quartile value within the rest of the samples. With the vast majority of genes having counts below 40 while some genes dominate the sample at 1,000,000 reads, we are concerned about what effects this phenomenon has on downstream analysis.

For comparison, the largest gap between any two consecutive, within sample, raw counts in the B. subtilis data set is only two-fold (footnote 2).

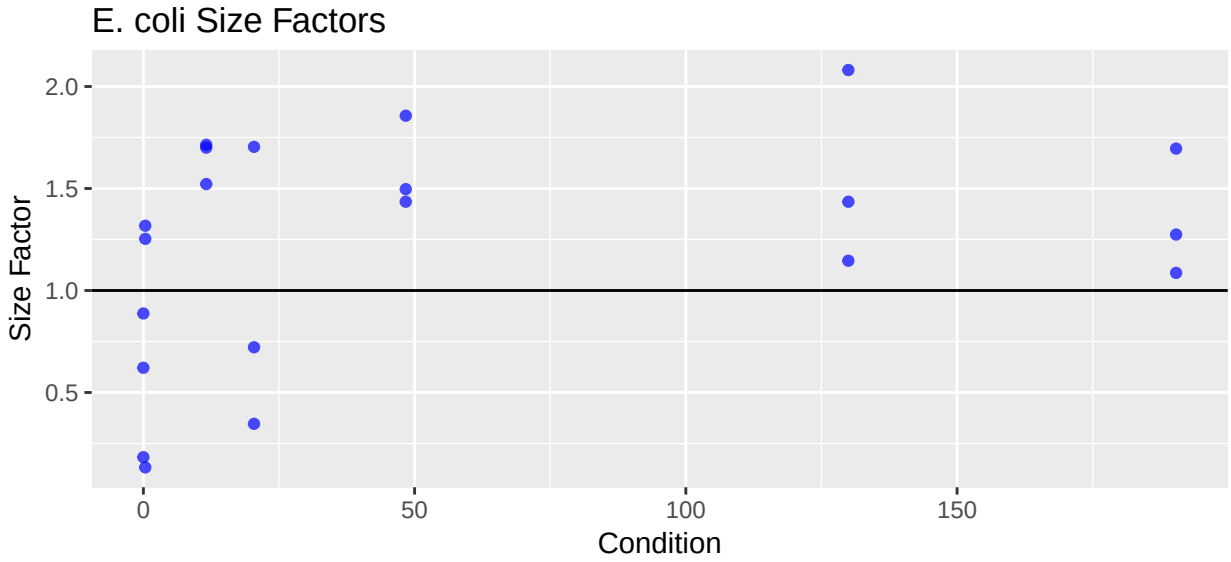
3. Highly Expressed Genes in E. Coli and B. subtilis Data Sets

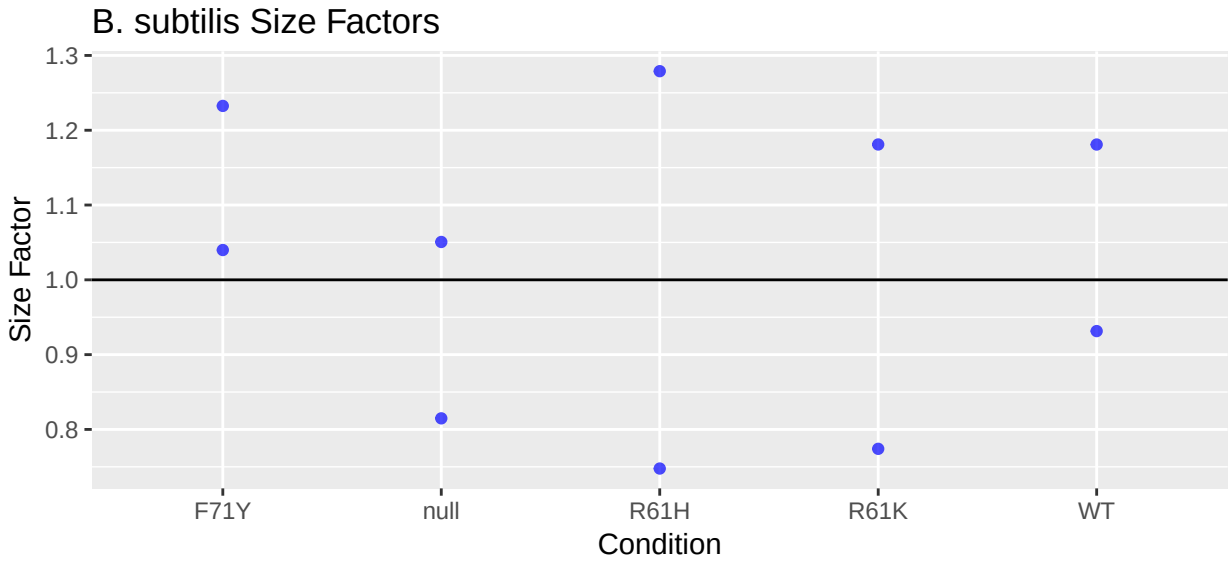
In the E. Coli data set, IGR smpB-intA has the highest read count in every sample, with between 708,292 and 1,597,980 counts per sample. IGR smpB-intA alone accounts for between 26% and 53% of the total reads for each sample. Within the set of IGR reads, IGR smpB-intA's reads alone constitute 85% to 98% of all sample-wise reads.

In the B. subtilis data set, there is also one gene which has the highest read count in every sample. This gene is BSU39230_wapA with between 243,705 and 519,330 counts per sample. BSU39230_wapA accounts for only between 2% and 4% of the total reads for each sample.

IGR smpB-intA is not the only gene with a high read count in the E. coli data, as we recall from the boxplots above. There are six genes whose counts exceed 100,000 and dominate the samples. These are CDS cspE, CDS rmf, IGR smpB-intA, ncRNA rna106, ncRNA rna 69, and ncRNA rna98. Defining our RpoS regulon as genes which are differentially expressed between the lowest and highest conditions, 0% (knockout) and 190.38%, all but CDS rmf are found in the regulon we would be considering for profile analysis (similar to the clustering analysis done in "Hierarchical expression of genes controlled by the Bacillus subtilis global regulatory protein CodY").

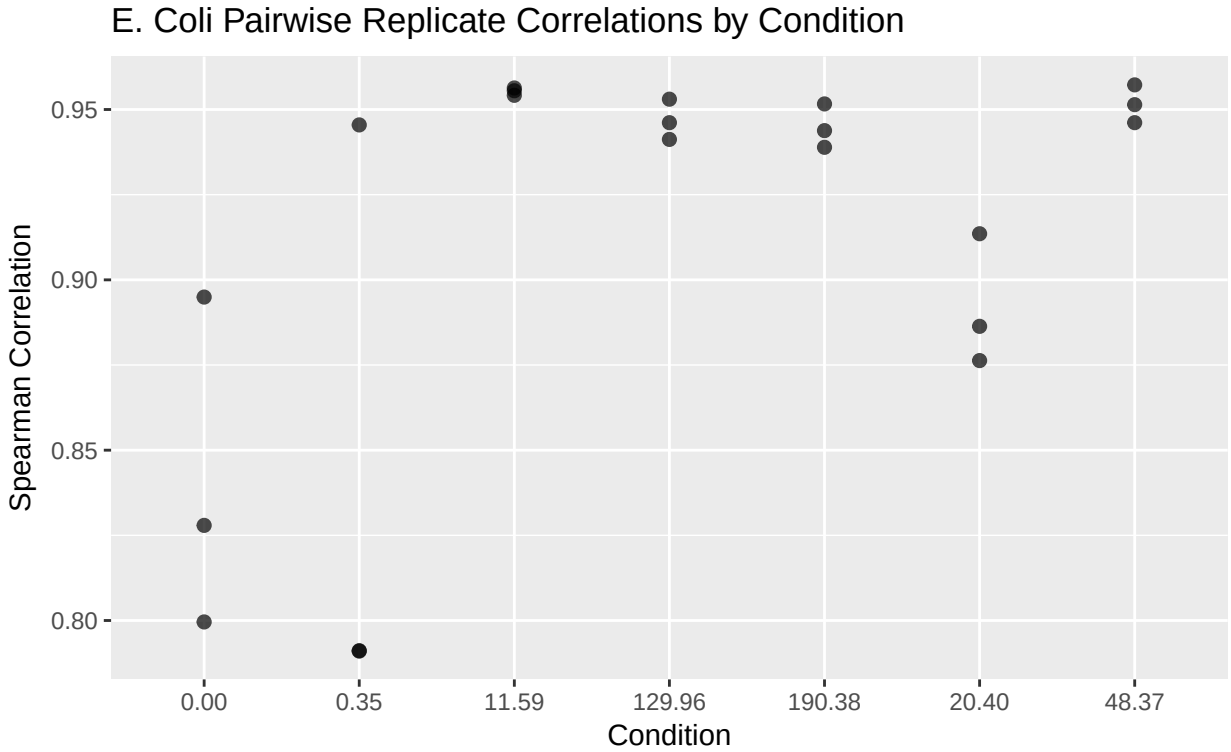
4. Size Factors





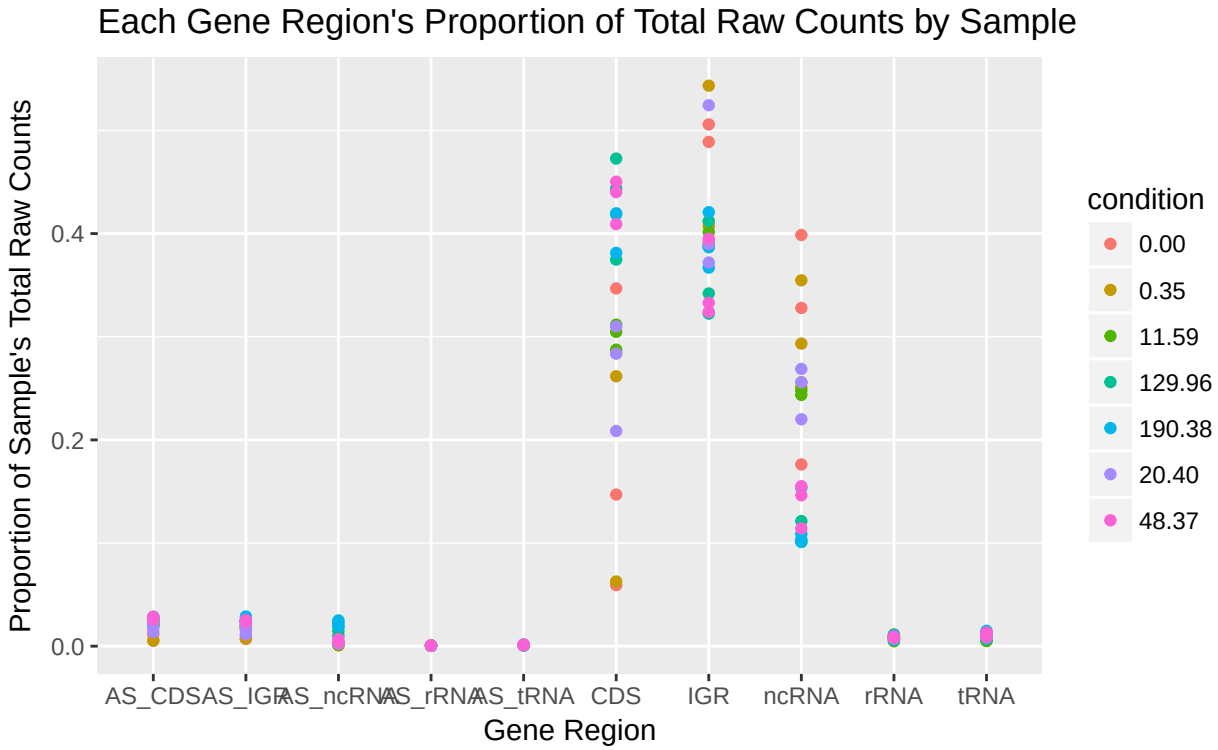
In the above plots, the size factors for each sample produced by DESeq2 are shown for E. coli and for B. subtilis. On the x-axis we see the conditions and on the y-axis we have the size factor; each condition having a size factor in blue for each of its replicates (3 replicates in the E. coli data and two replicates in the B. subtilis data). The size factors produced for the E. coli data are much more spread, ranging from 0.13 to 2.08, than the size factors produced for the B. subtilis data. This phenomenon makes sense recalling the zoomed-in read count boxplots from part 2a.

5. Correlations Between Replicates



Each point on the above plot represents the pairwise Spearman correlation between two of three replicates for each condition. The Spearman correlations range between 0.791 and 0.957. For comparison, Spearman correlations between the two replicates at each condition for B. subtilis were between 0.973 and 0.992.

6. Proportion of Reads From Each Feature in the E. coli Data Set



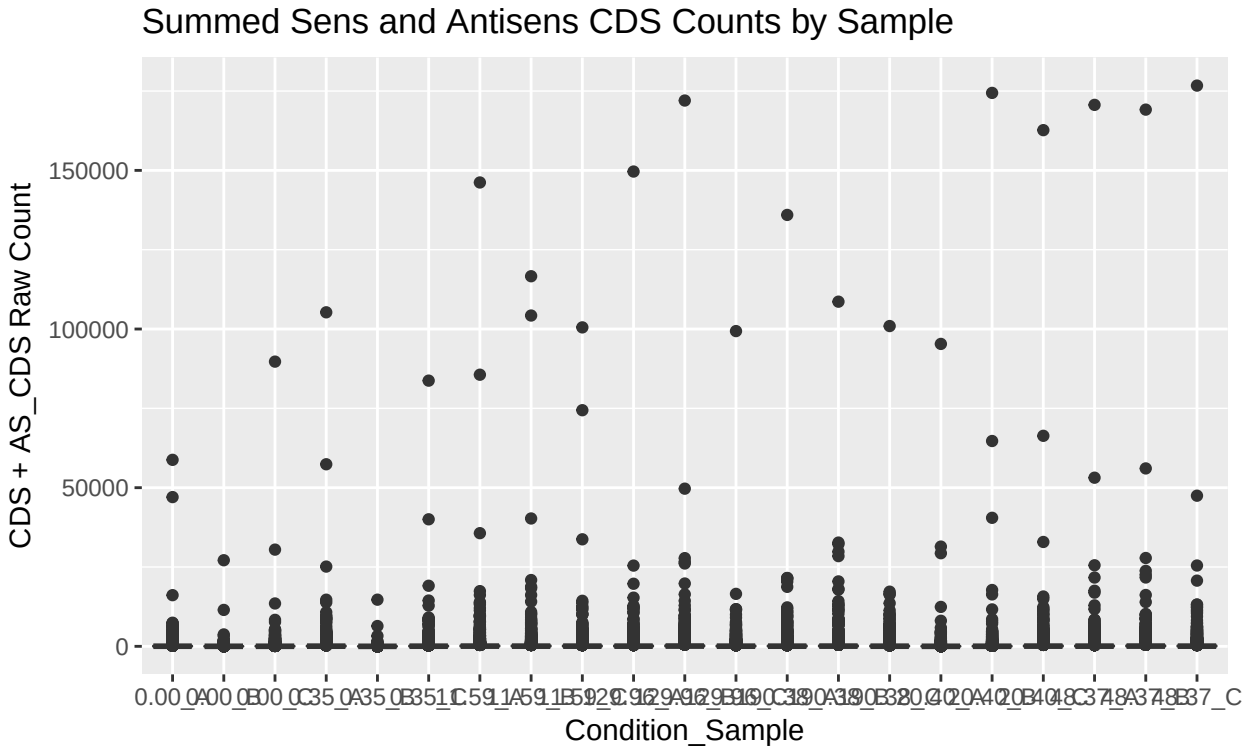
The proportions in the above plot are calculated in the following way. Summing each sample's raw read counts, we obtain the total raw read count for each sample. Then, within each sample, we group by gene feature to obtain a total raw read count for each feature within each sample. We then take the latter over the total raw read count for each sample to calculate the proportion of reads for each feature within each sample. As evidenced by the above plot, the proportion of per-sample gene feature reads is not consistent within each gene feature. CDS is perhaps the most striking gene feature, with proportions varying from 0.059 to 0.473 depending on the sample. IGR and ncRNA fare better, but still show a difference of approximately 0.2 and 0.3, respectively, between their lowest and highest proportions.

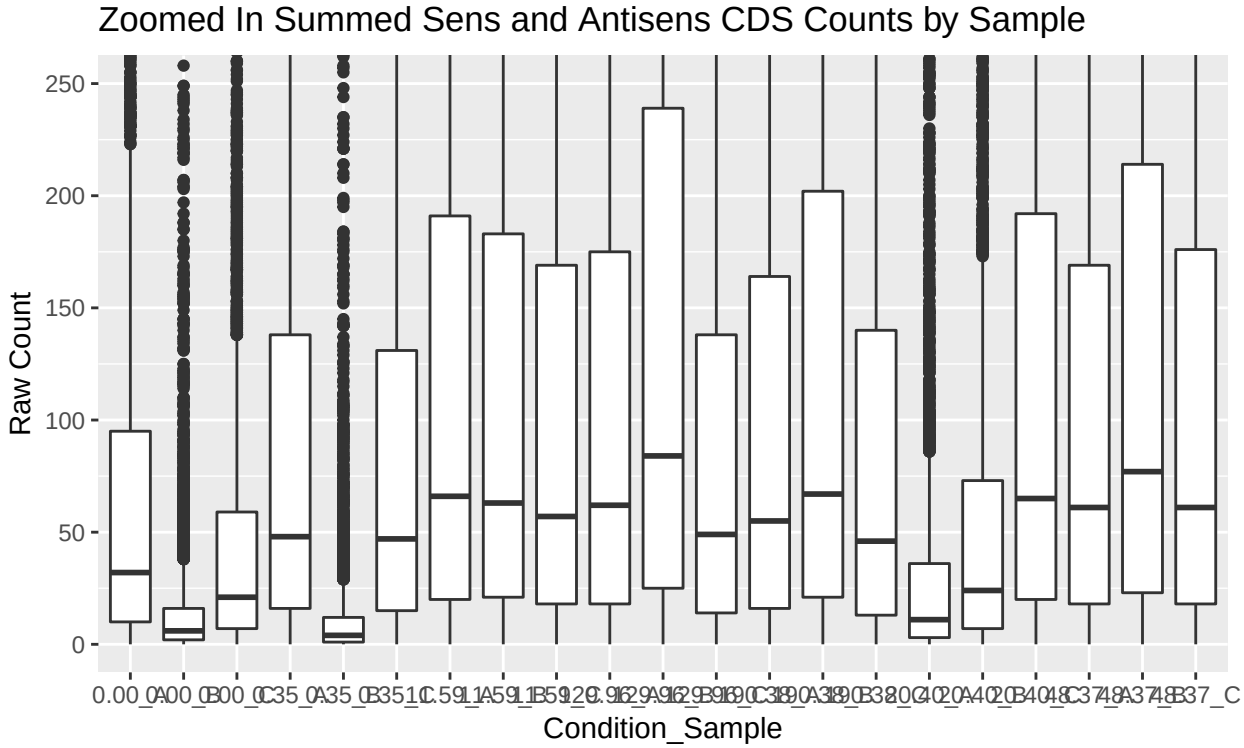
Interestingly, for the samples with single digit percentages of CDS, the bulk of those samples are comprised of IGR and ncRNA. For example, 5.9% of the reads in sample 0.00_B were CDS, while 50.6% of its reads were IGR and 39.9% of its reads were ncRNA.

Footnotes

Footnote 1:

We understand that the summed sens and antisens CDS counts are a more appropriate direct comparison to the Bacillus subtilis data. When creating box plots of the raw counts by sample for just this portion of the data (below), we notice that the counts are not as high as within the entire data set (no raw count exceeds 200,000). We also notice that the medians and thrid quartiles are higher, so the lower 75% of the AS_CDS+CDS data has higher read counts than the lower 75% of all gene features, but the lower 75% within AS_CDS+CDS still does not exceed 200 for nearly all samples. There still exist substantial differences between the median and third quartile read counts of each sample, appearing to vary even more greatly than in the combined data.





Footnote 2:

This two-fold difference is seen in R61H2 between the genes BSU01120_fusA_elongation_factor_G and BSU39230_wapA_cell_wall-associated_protein.

EXTRA:

Unconstraining the y-axis of the raw counts boxplot from above, we notice that there exist wide gaps between raw counts. For instance,

For comparision, we looked at the raw counts in "Hierarchical expression of genes controlled by the Bacillus subtilis global regulatory protein CodY" (footnote 1). Below, we see boxplots of the Bacillus subtilis raw counts by sample, first with unconstrained y-axis, then with a constrained y-axis from 0 to 3,000. As evidenced from the plots, the highest raw counts do not exceed 520,000 and 50% of the counts for each sample are between roughly 60 and 2,500. Thus although the sequencing depth in Bacillus subtilis is lower than the E. coli data, the medians and thrid quartiles of raw counts are higher. Furthermore, there is not such a drastic difference between the medians and third quartiles across samples as we see in the E. coli data. Finally, it is worth noting that the largest gap between any two consecutive, within sample, raw counts is only two-fold (footnote 2); much smaller than in the E. coli data.

Certainly, these differences between the E. coli and Bacillus subtilis datasets could be an artifact of biology, and different roles of genes in their respective genomes. Yet it is still worth wondering why the E. coli raw counts show such a majority of low raw counts in each sample, with a minority of very high counts.

In the E. coli data, the great difference between third quartiles and maximum read counts per sample does not change appreciably across gene features, and we continue to see large gaps between raw counts by gene whether or not we look at CDS, IGR, ncRNA, etc. individually. We also notice that, generally, genes with large read counts are large across all the samples.

In the E. coli data, there are

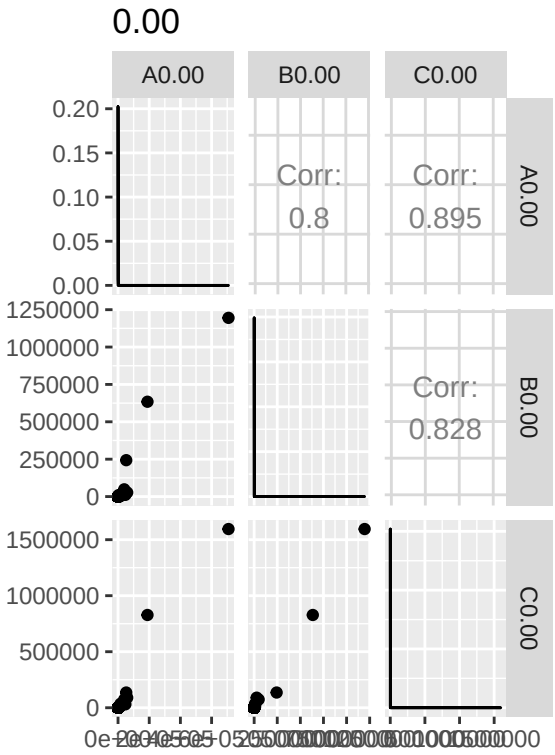
IGR smpB-intA, mentioned above, has nearly all samples with read counts around 1,000,000. This gene's reads alone constitutes 85% - 98% of all sample-wise IGR reads, and 28% - 53% of reads across all gene features.

The gene ncRNA rna69, whose total read count per sample ranges from around 300,000 to 800,000, makes up 11% - 75% (mean 50%) of all ncRNA reads, and 1% to 26% (mean 12%) of reads across all gene features.

The CDS genes, cspE and rmf, make up lower proportions of their own features counts and the counts across all gene features. The raw reads of rmf account for 8% to 24% of sample-wise CDS total reads, and 1% to 7% of reads across all gene features. The raw reads of cspE account for 1% to 8% of sample-wise CDS total counts, and 0.4% to 2% of reads across all gene features.

Inter-Replicate Correlations by Condition

The Pearson intra-replicate correlations by sample range between 0.9742 to 0.9976, likely influenced by the high count genes. Below are pairs plots to show the most spread correlations (knockout condition) and a



more typical distributions of correlations (129.96 condition).