Quality Control Report

by

Barry Digby

Orbsen Therapeutics
Dangan Heights, Galway Business Park, Galway, H91 EFD0

Supervisors: Dr. Pilib Ó Broin

In collaboration with: Elaine Cullen

Contents

1 Quality Control of Samples

QC of the donor samples is carried out to adress the following questions:

- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiments design?
- What are the major sources of variation in the dataset?

To explore the similarity of the samples, sample-level QC using Principal Component Analysis (PCA) and hierarchical clustering methods were employed. Sample-level QC illustrates how well the replicates cluster together, as well as, observe whether our experimental condition represents the major source of variation in the data. Performing sample-level QC can also identify any sample outliers, which may need to be explored further to determine whether they need to be removed prior to DE analysis.

1.1 Sample to Sample distance Heatmap

The euclidean distances were extracted from the regularized logarithm data and implemented to create a heatmap of similarity between the samples. As can be seen in Figure 2.1, the samples cluster together based on sample condition: Control vs. CYTO (treated). This clustering was to be expected from the samples. Also of note is the fact that each sample donor (H88 - H91) tend to cluster together along the rows of the heatmap which indicates that cell types within sample donors are more similar to each other than the cell types themselves. I would have expected Aria, Aria_S2, PA and TYTO_S2 to form clusters within each sample condition regardless of donor.

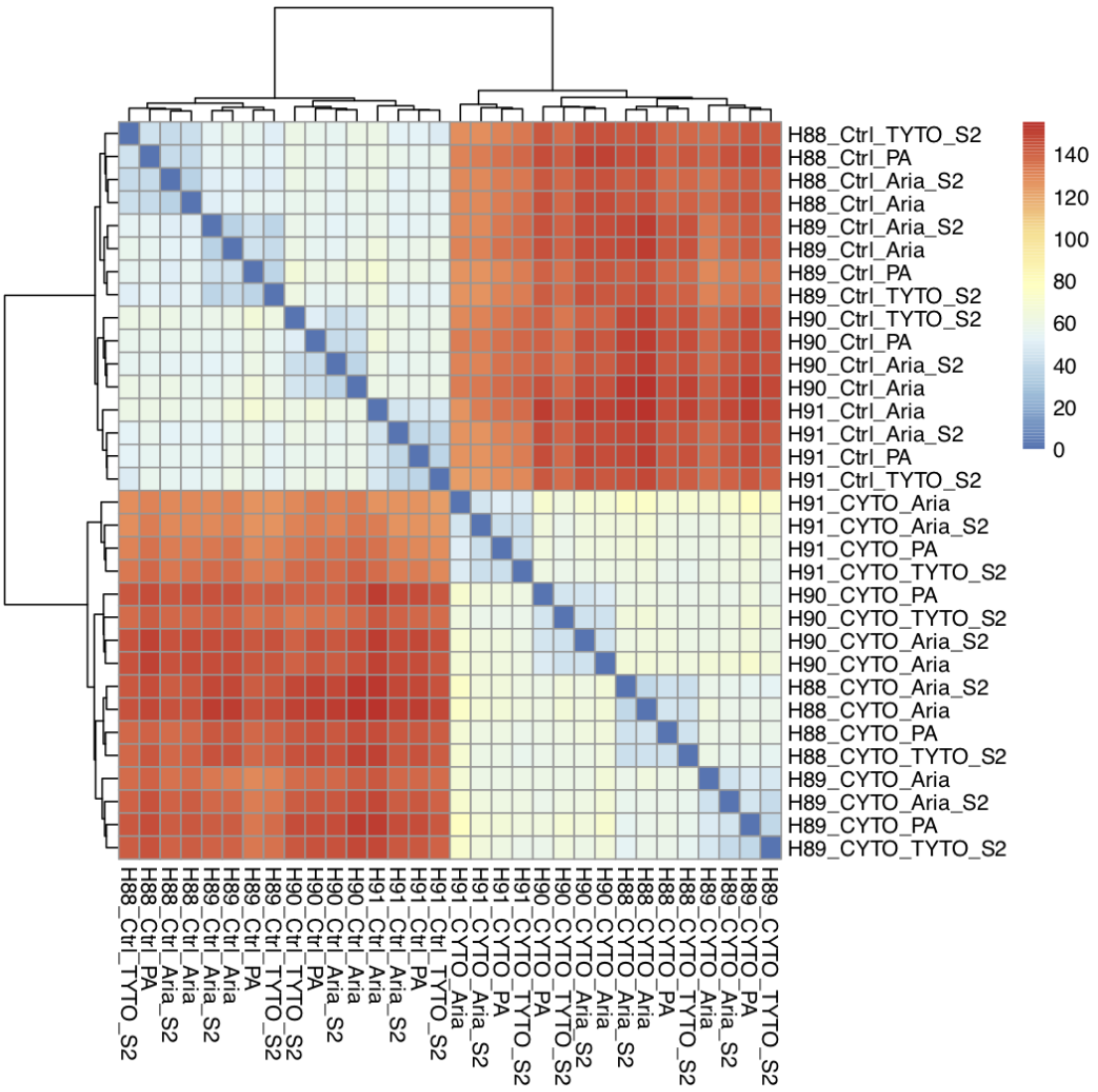

Figure 1.1: Sample to Sample Heatmap of all 32 samples.

1.2 Mapped reads per Sample

It is useful to inspect how many reads were PseudoAligned by Kallisto, as samples with significantly higher or lower reads aligned can effect the statistical power when conducting Differential Gene Expression analysis. Figure 2.2 shows a bar chart of each sample, displaying a minimum value of 30.99, a mean of 34.08, median of 34.00 and maximum value of 36.69 million reads, respectively.

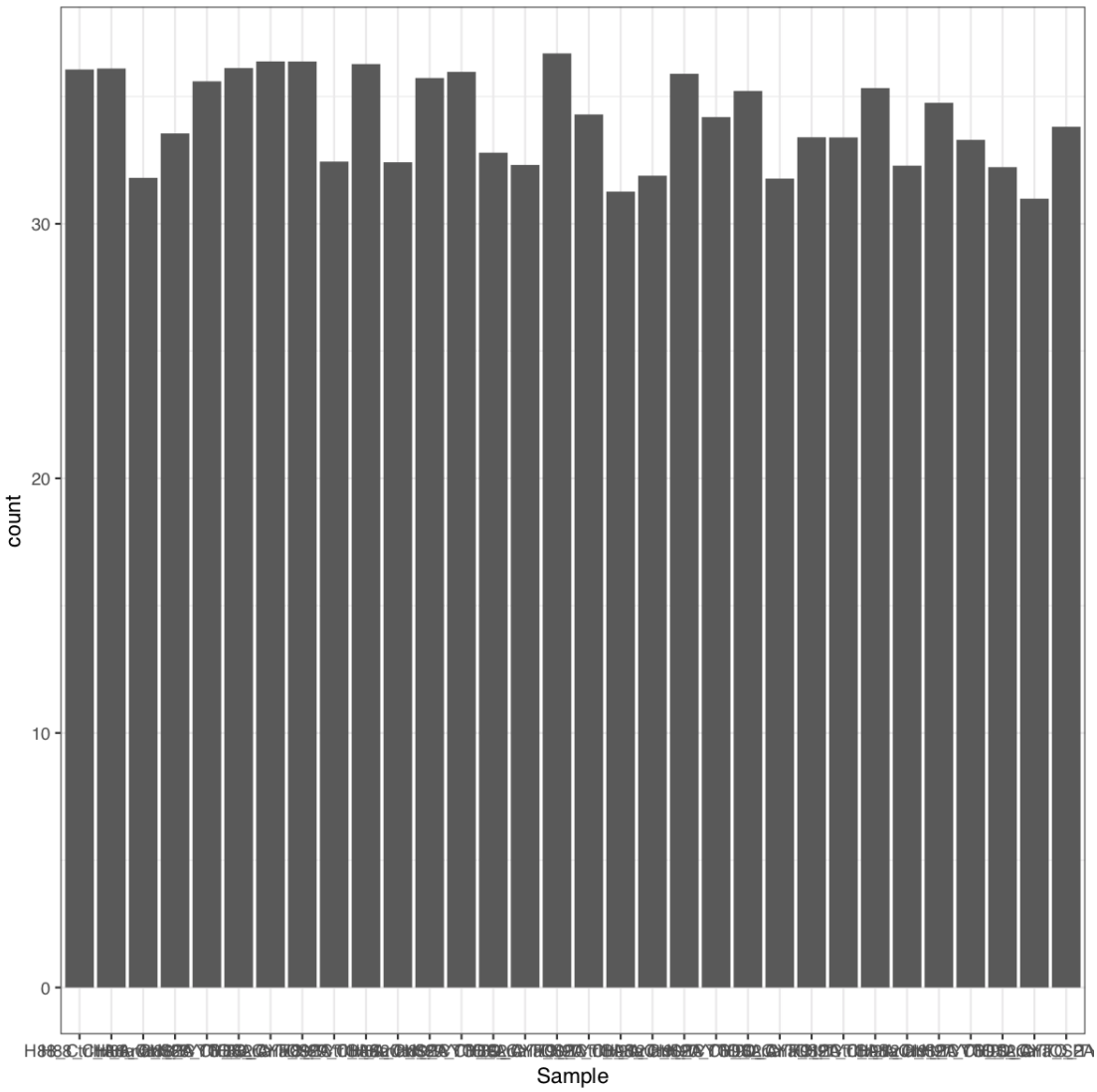

Figure 1.2: Mapped reads by Kallisto.

1.3 PCA

Due to the large nature of expression data, in order to accurately plot or compare sample groups, the data has to be combined in lower dimensional space. Methods include principal component analysis (PCA) which attempts to transform a large set of variables into a smaller one that still contains most of the information in the large set. Figure 2.3 shows the scree plot for the samples. A scree plot is a line

plot of the eigenvalues of principal components in an analysis, showing the largest sources of variation. As is evident in Figure 2.3, PC1 has the largest value and accounts for the largest source of variability in the samples. To figure out what the driving factor for PC1 is, we can use a PCA plot to contrast between sample donors, condition, cell type etc and see if the samples form clusters accordingly.

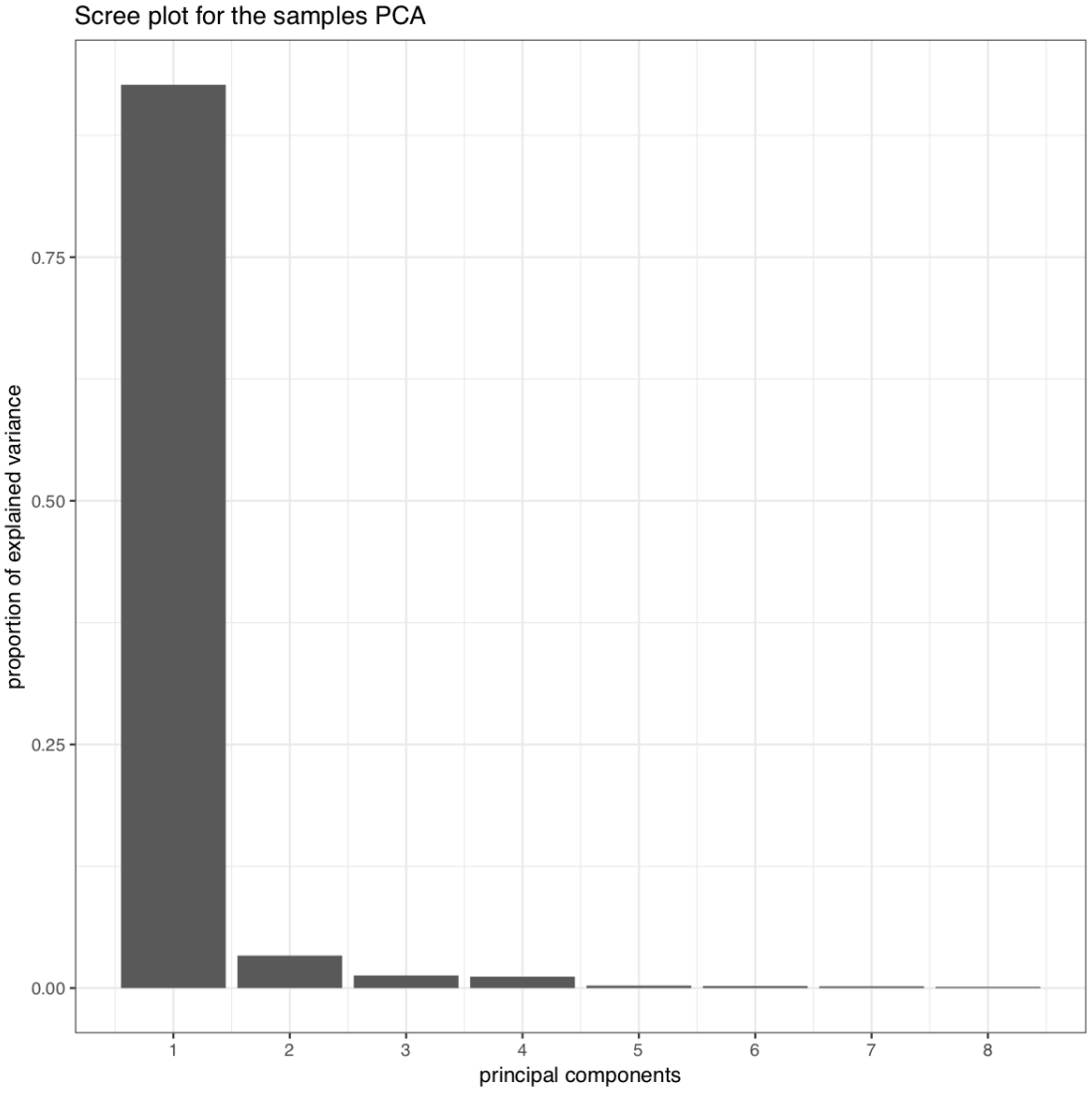

Figure 1.3: Scree plot for samples.

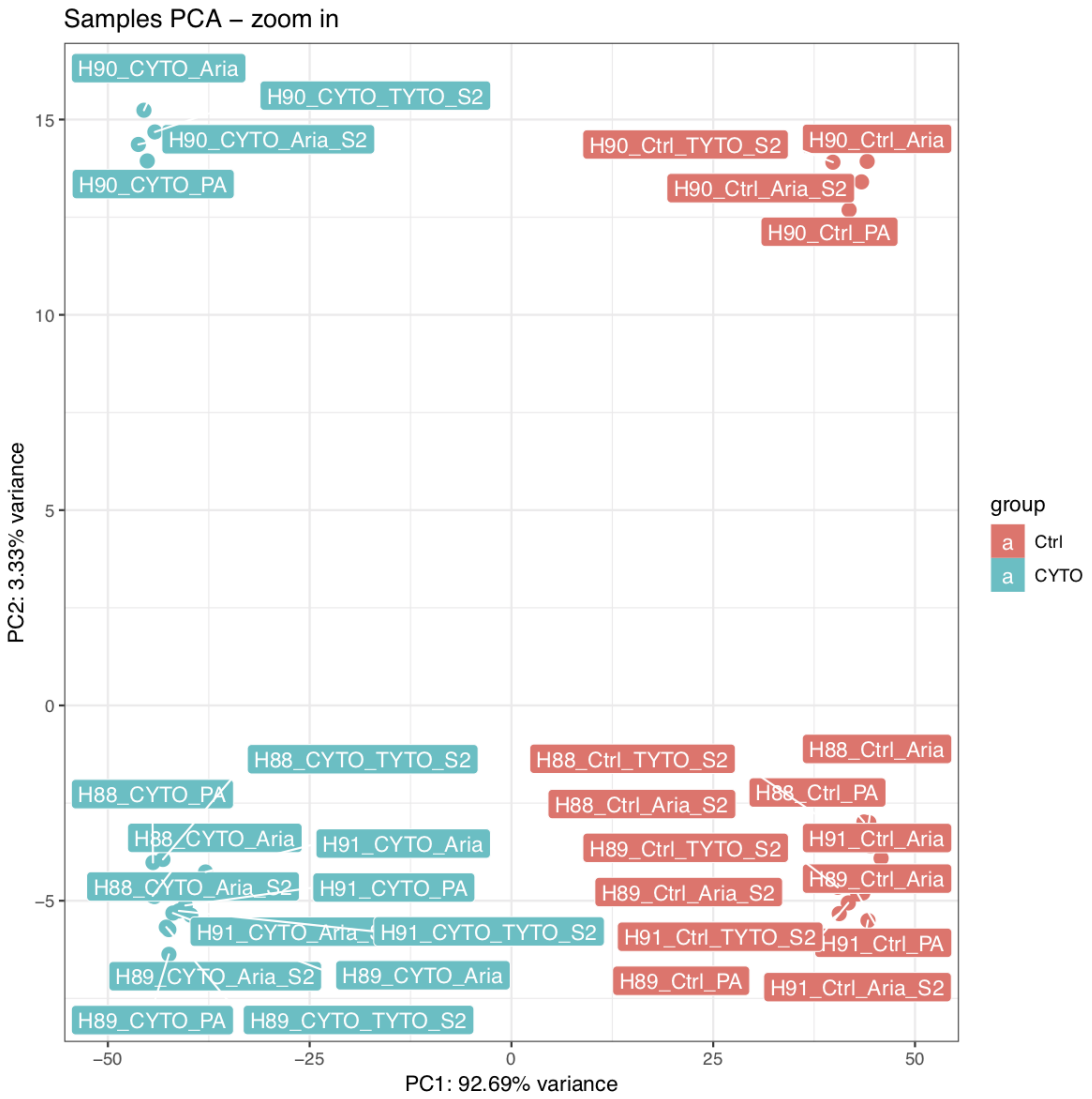

Figure 1.4: PCA for the samples.

In Figure 2.4, the samples cluster on opposite sides of PC1 according to sample condition. This is in concordance with the sample to sample heatmap shown in section 2.1. Interestingly, samples from Donor H90 form clusters above their respective sample condition, indicating a 3.33% source of variance in PC2. The PCA plots indicate that the most meaningful results will be derived from comparing cell types across sample conditions e.g: Ctrl_Aria vs CYTO_Aria, and not between cell conditions e.g Ctrl_Aria vs Ctrl_AriaS2 as the researcher had requested.

1.4 PCA Loadings

As aforementioned in section 2.3, the samples from donor H90 clustered away from its counterparts. To investigate this further, we can inspect the loadings on PC2 (genes driving this 3.33% variation) by plotting box plots of the top 3 variable genes and attempt to draw conclusions from the data.

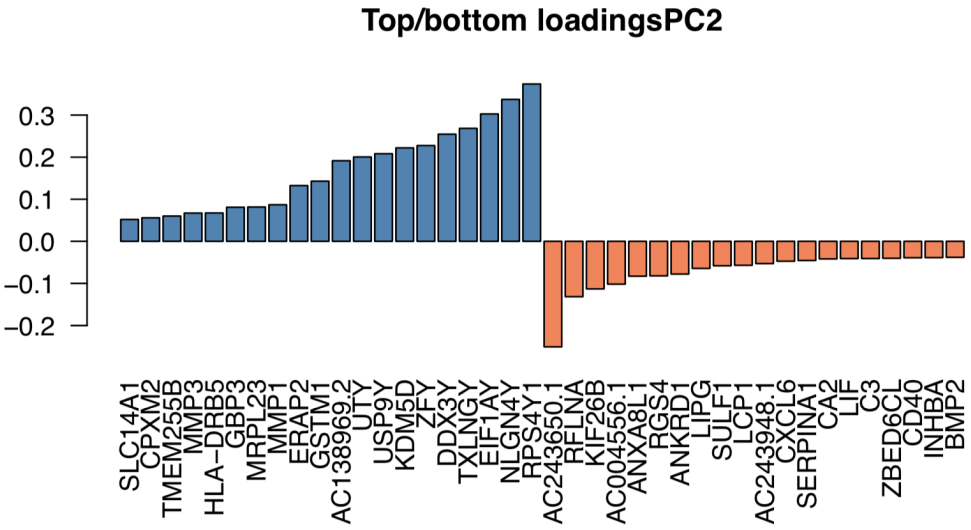

Figure 1.5: Loadings on PC2 driving the variation of sample donor H90.

1.4.1 Highly Expressed genes in sample donor H90

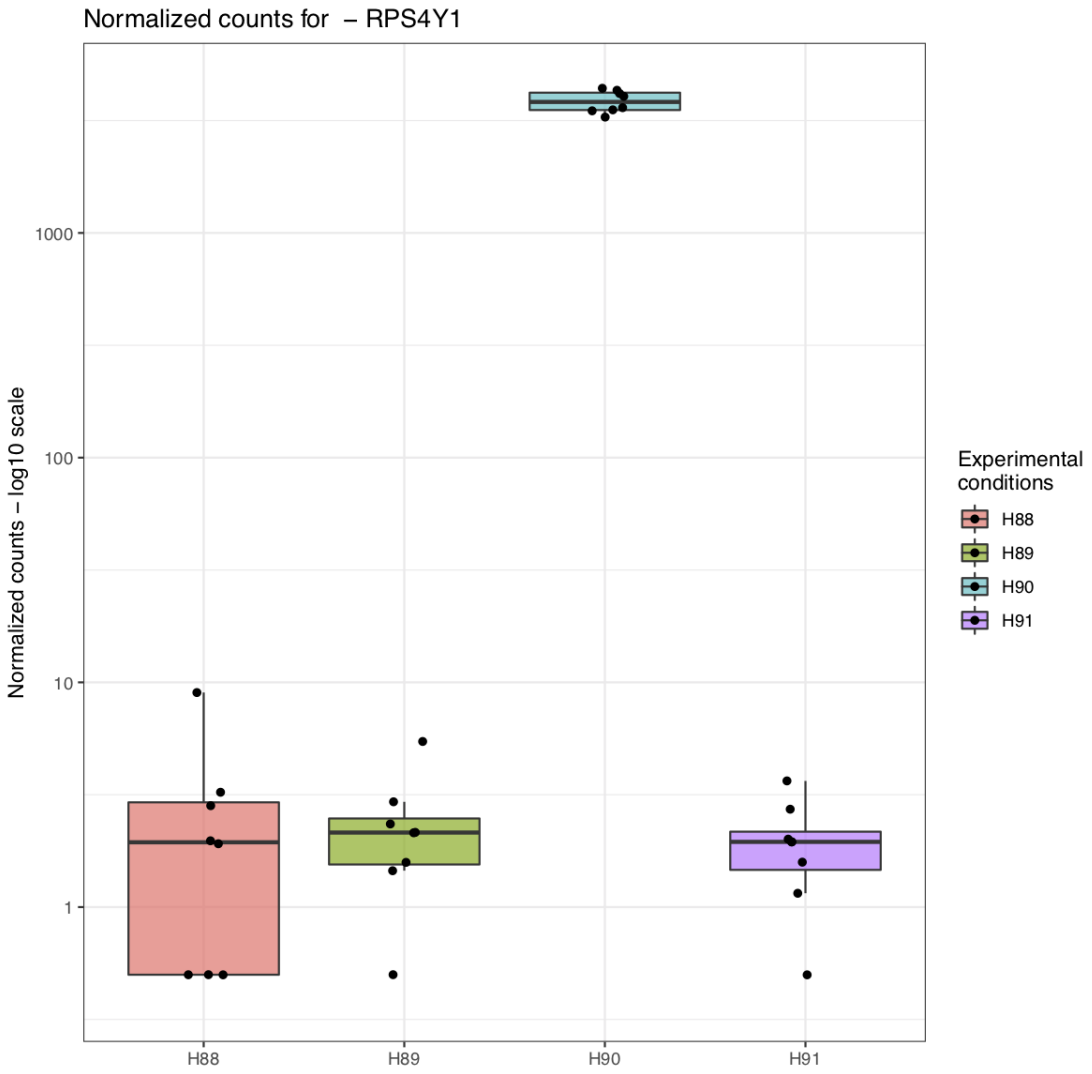

Figure 1.6: Boxplot of the RPS4Y1 gene across sample donors.

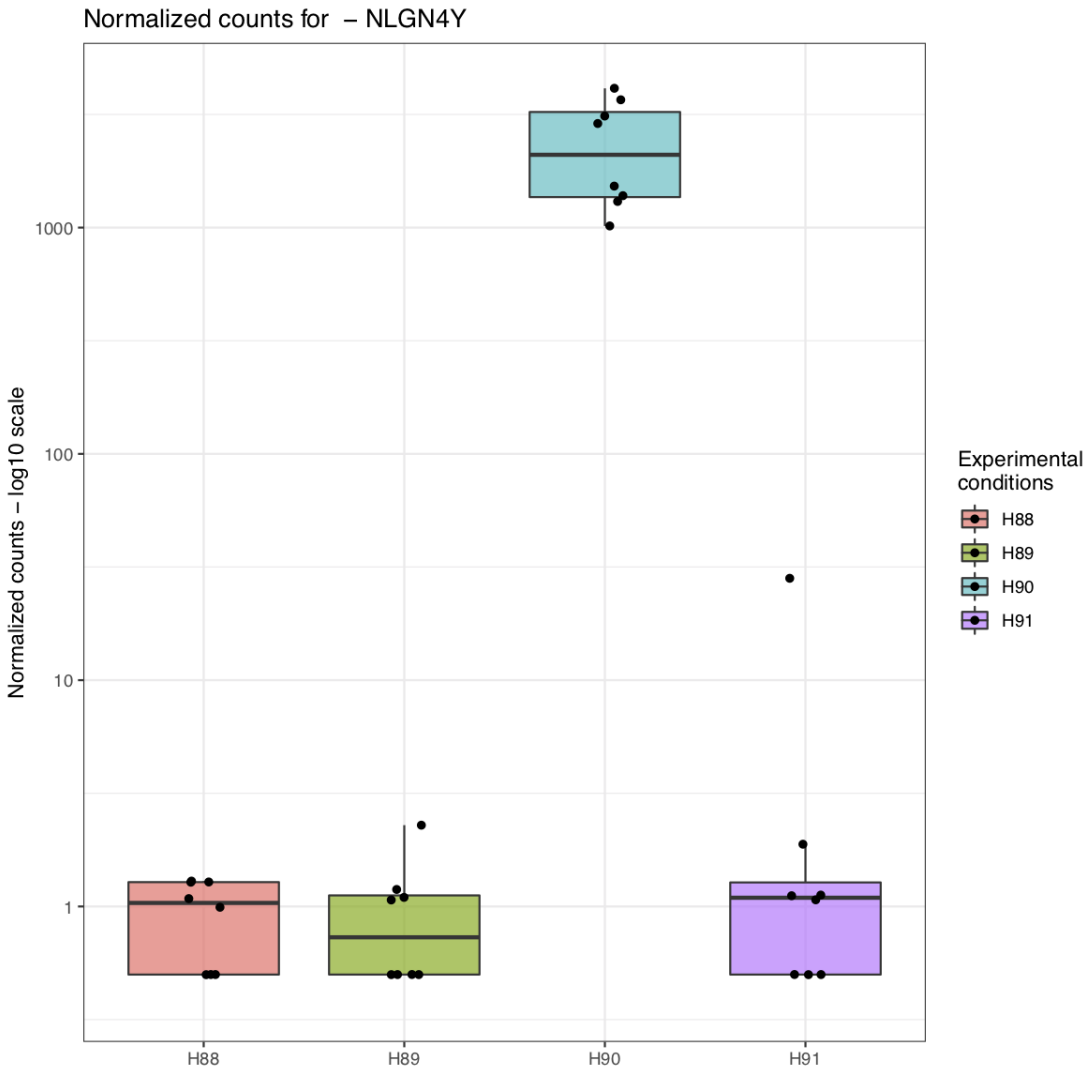

Figure 1.7: Boxplot of the NLGN4Y gene across sample donors.

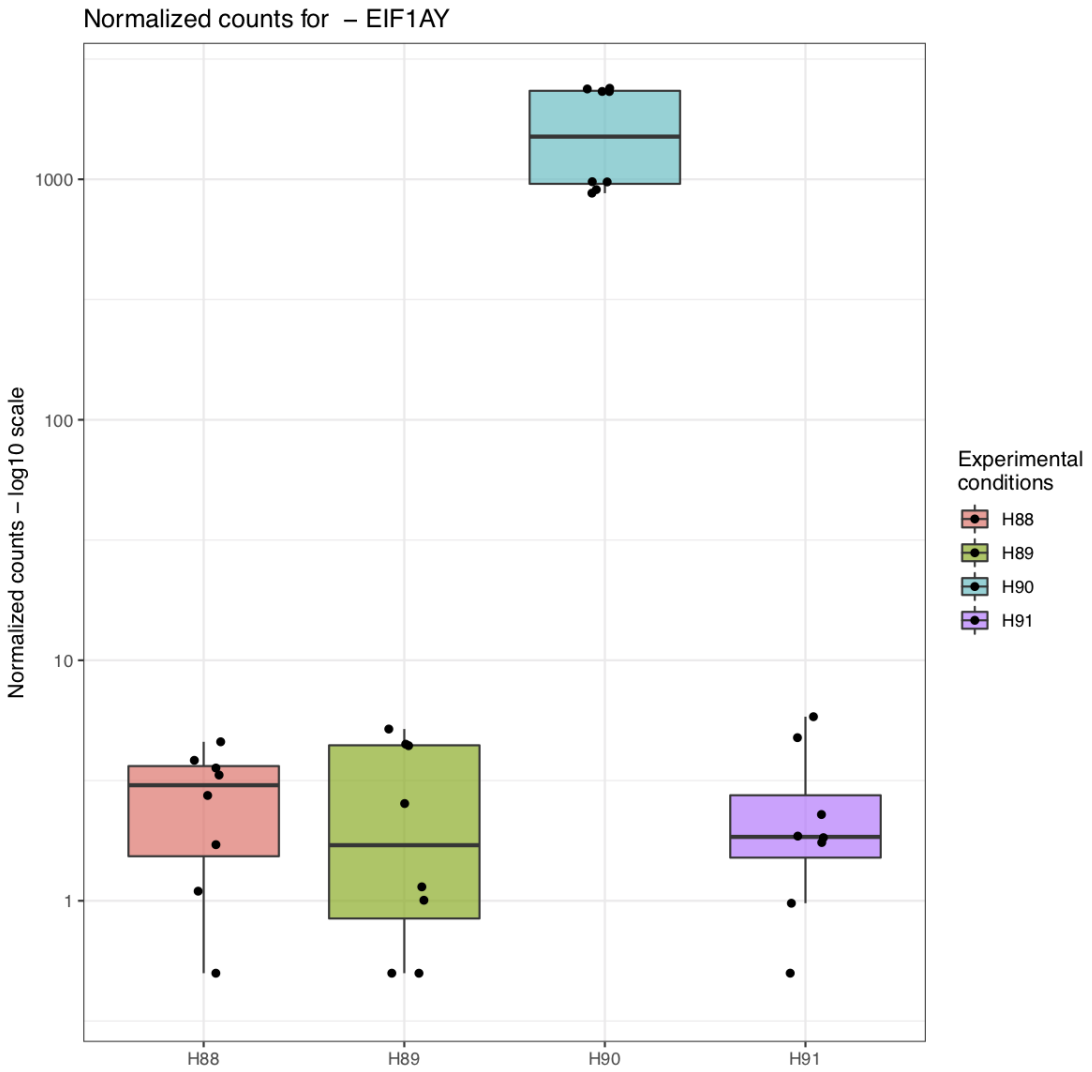

Figure 1.8: Boxplot of the EIF1AY gene across sample donors.

From the boxplots above, we can see that these genes are highly expressed in sample donor H90 compared to the other samples. Fortunately, DESEQ2 accounts for outlier genes in samples when constructing the DDS object, and these outliers should not effect Differential Gene Expression analysis. However as a measure of precaution when conducting pairwise analysis of cell types within sample condition, I will remove these samples and see if it effects the number of genes differentially expressed.

1.4.2 Lowly Expressed genes in sample H90

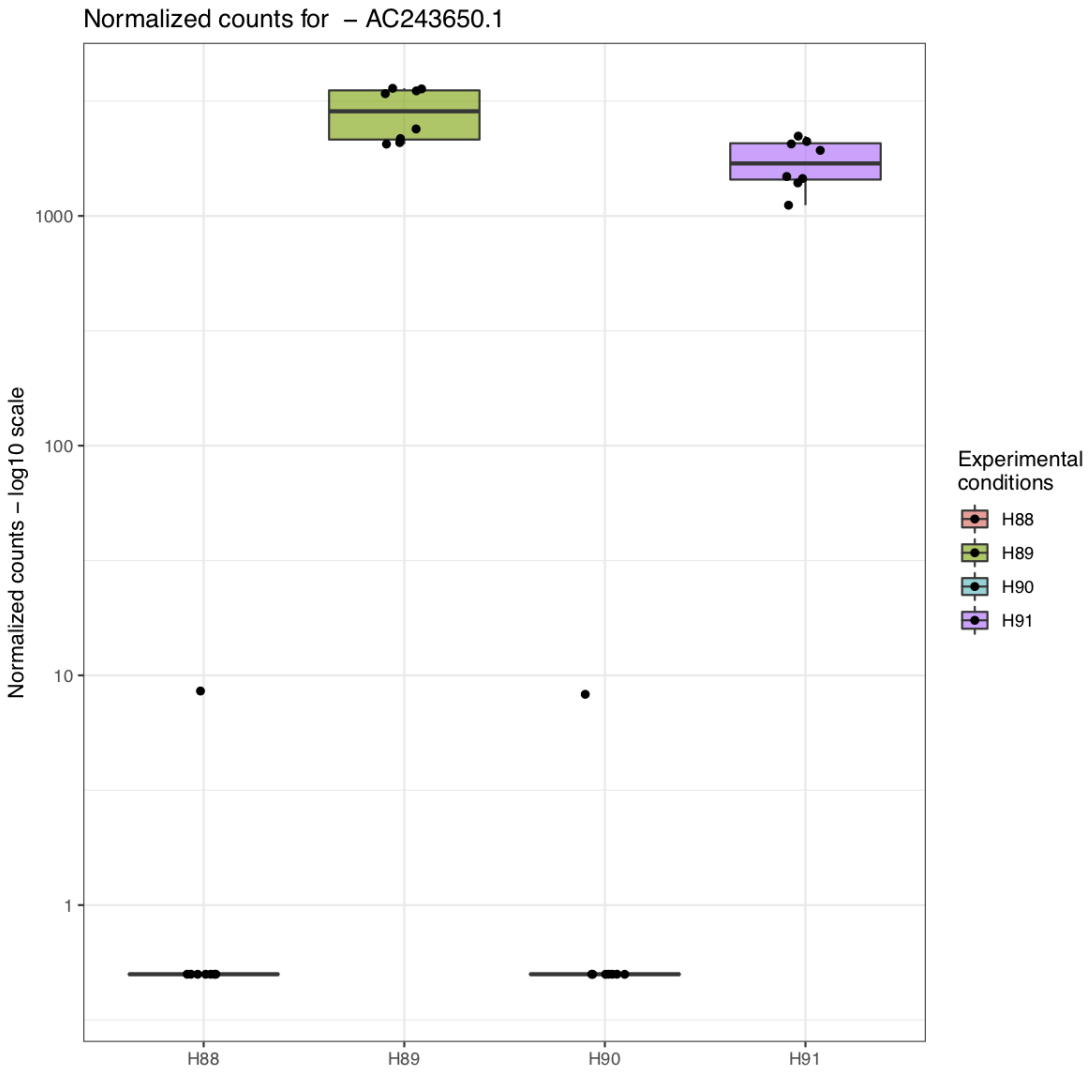

Figure 1.9: Boxplot of the AC243650.1 gene across sample donors.

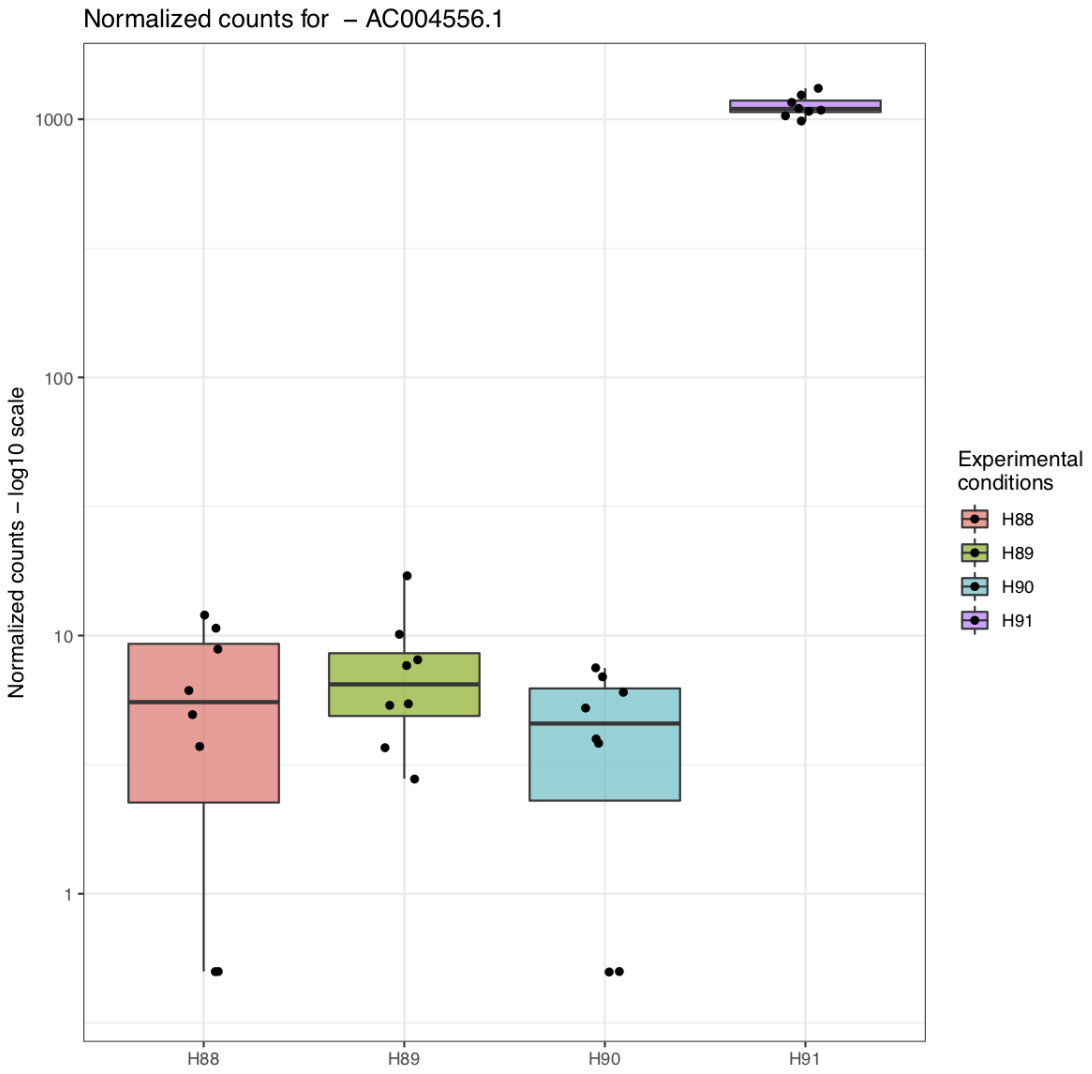

Figure 1.10: Boxplot of the AC004556.1 gene across sample donors.

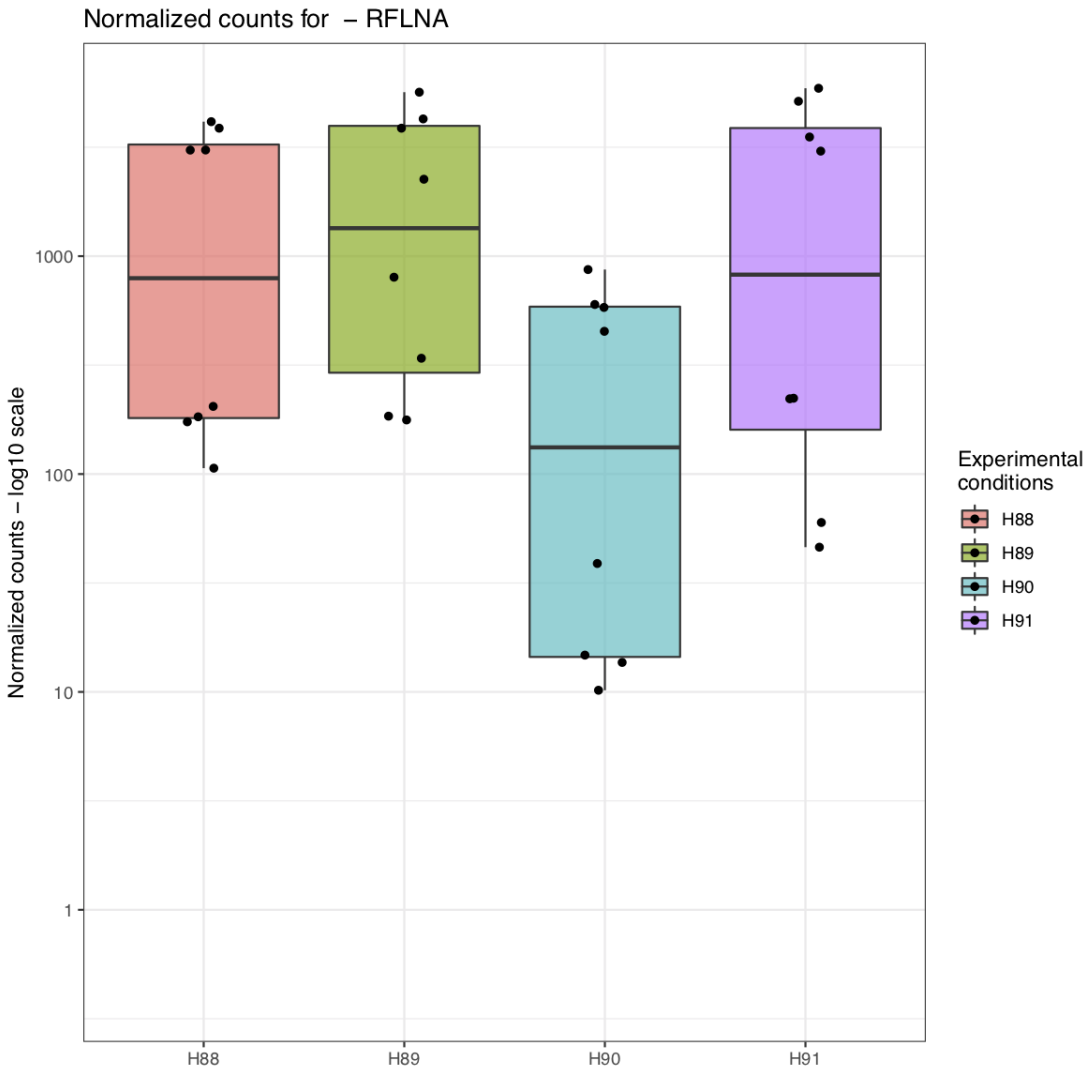

Figure 1.11: Boxplot of the RFLNA gene across sample donors.

The results of the top 3 loadings negatively effecting PC2 show mixed results across samples which means we can concur that these sources of variation are not driven by a defective sample donor H90. We can conclude that the main source of variation in PC2 is a handful of genes that are highly expressed in sample H90, and was picked up by the high sensitivity of PCA analysis.

1.5 Hypothesis

For this analysis I will be testing the *null hypothesis* that there is no effect of cell type on the samples and that observed differences between cell types was merely caused by experimental variability. The results of this test will be reported in *pvalues*, the probability that a fold change would be seen under the effect of cell types (rejecting the null hypothesis). Further statistical filtering will be applied (Benjamini & Hochberg) to calculate the fraction of false positives amongst those rejecting the null hypothesis.