Supplementary material for "Visualization methods for RNA-sequencing data analysis"

Lindsay Rutter

March 6, 2018

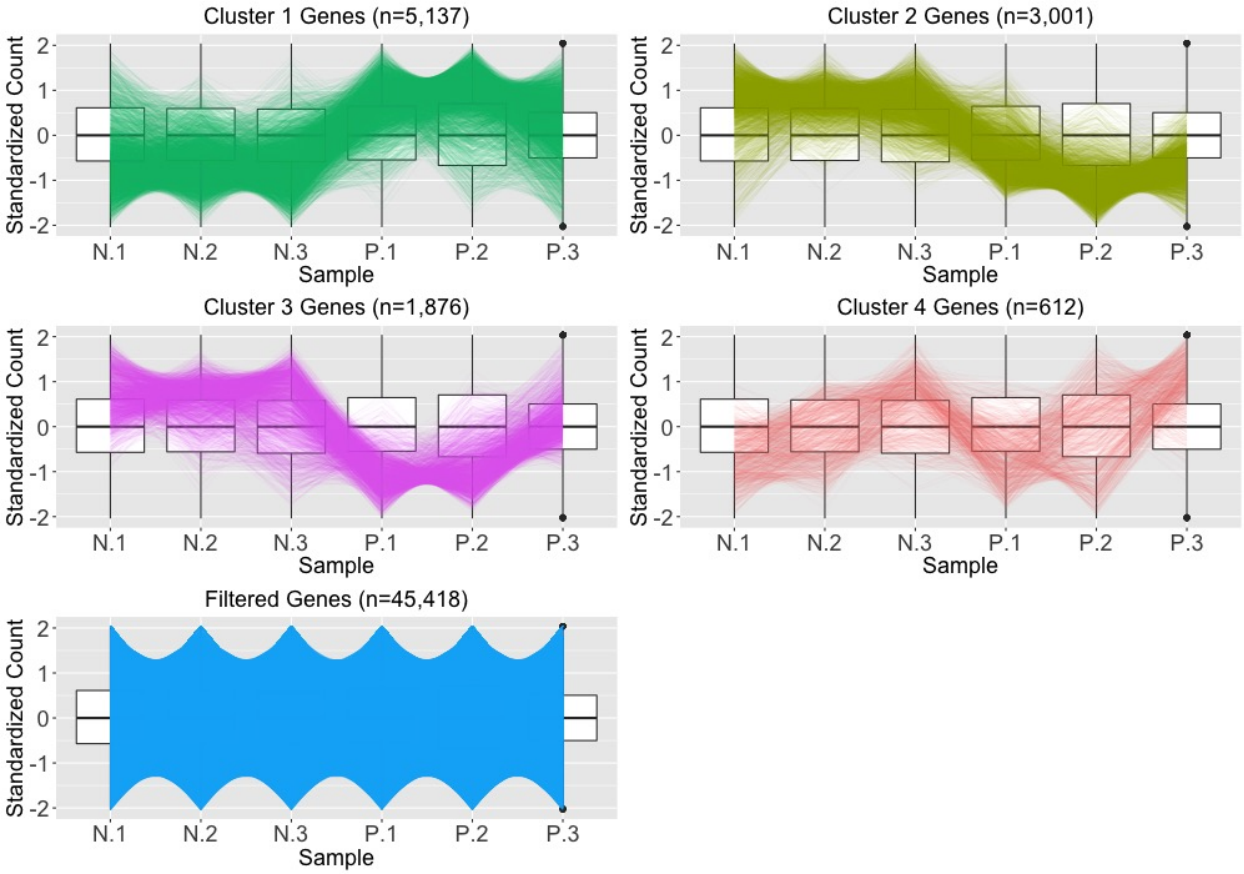

Figure 1: Example application of parallel coordinate plots using the iron-metabolism soybean dataset. We filtered genes with low means and/or variance, performed a hierarchical clustering analysis with a cluster size of four, and visualized the results using parallel coordinate lines. Most non-filtered genes were in Clusters 1 and 2, which both showed overexpression in one treatment and underexpression in the other treatment. The genes in Cluster 4 mostly showed messy patterns with low signal to noise ratios. Interestingly, Cluster looked similar to Cluster 2 (large values for group N and small values for group P), except for unexpectedly large values for the third replicate of group P.

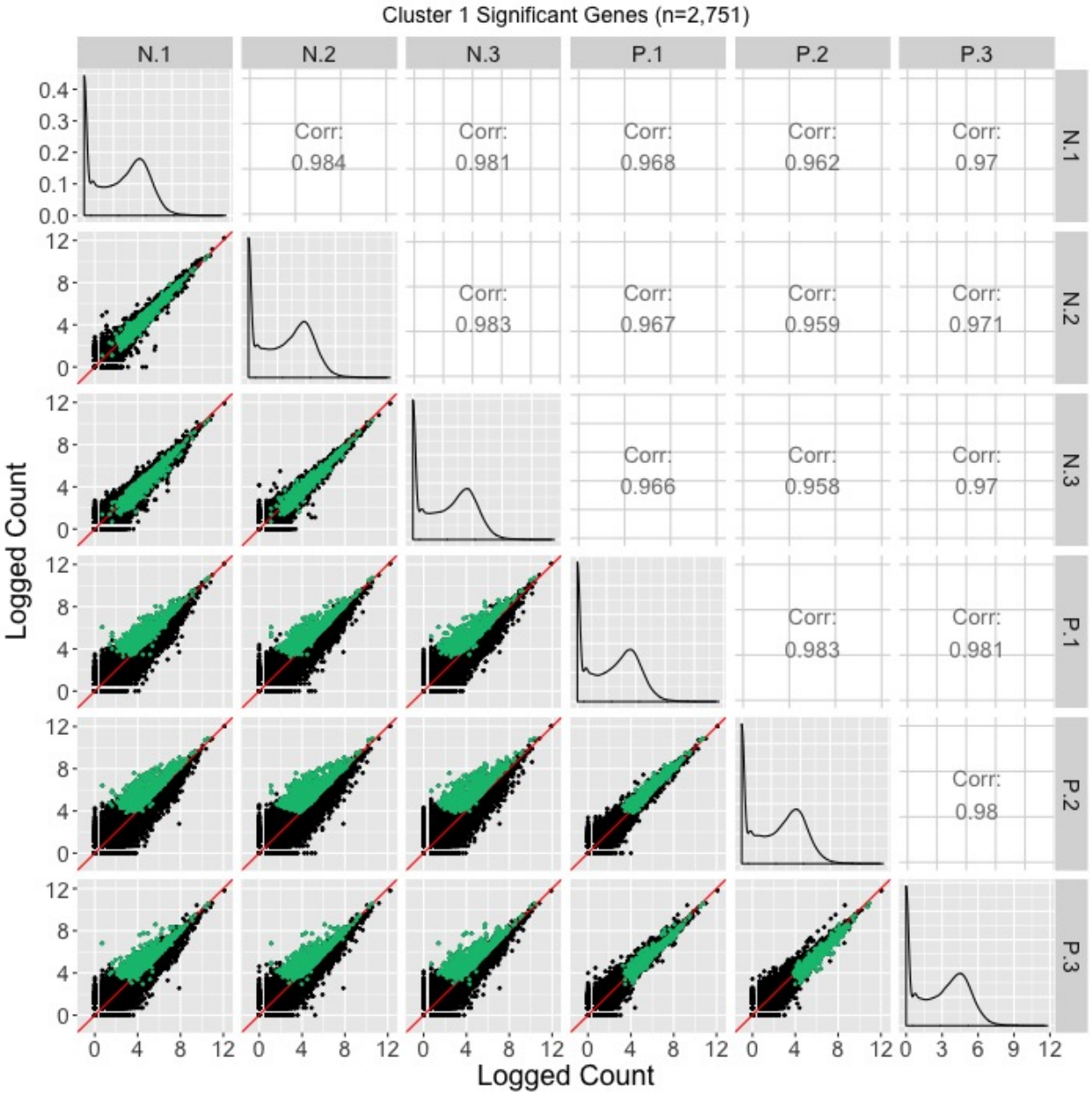

Figure 2: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2751 significant genes in Cluster 1 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in green over the scatterplot matrix. They follow the expected patterns of differential expression with most green points falling along the x=y line in the scatterplots between replicates, but deviating from the x=y line in the scatterplots between treatments. The deviation consistently demonstrates higher expression in the P group than in the N group. Hence, these green points seem to represent genes that were significantly overexpressed in the P group, which draws the same conclusion with what we derived using the parallel coordinate plots in Figure 2 of the paper. One difficulty with plotting such a large number of DEGs onto the scatterplot matrix is that overplotting can obscure our inability to determine how many DEGs are in a given location.

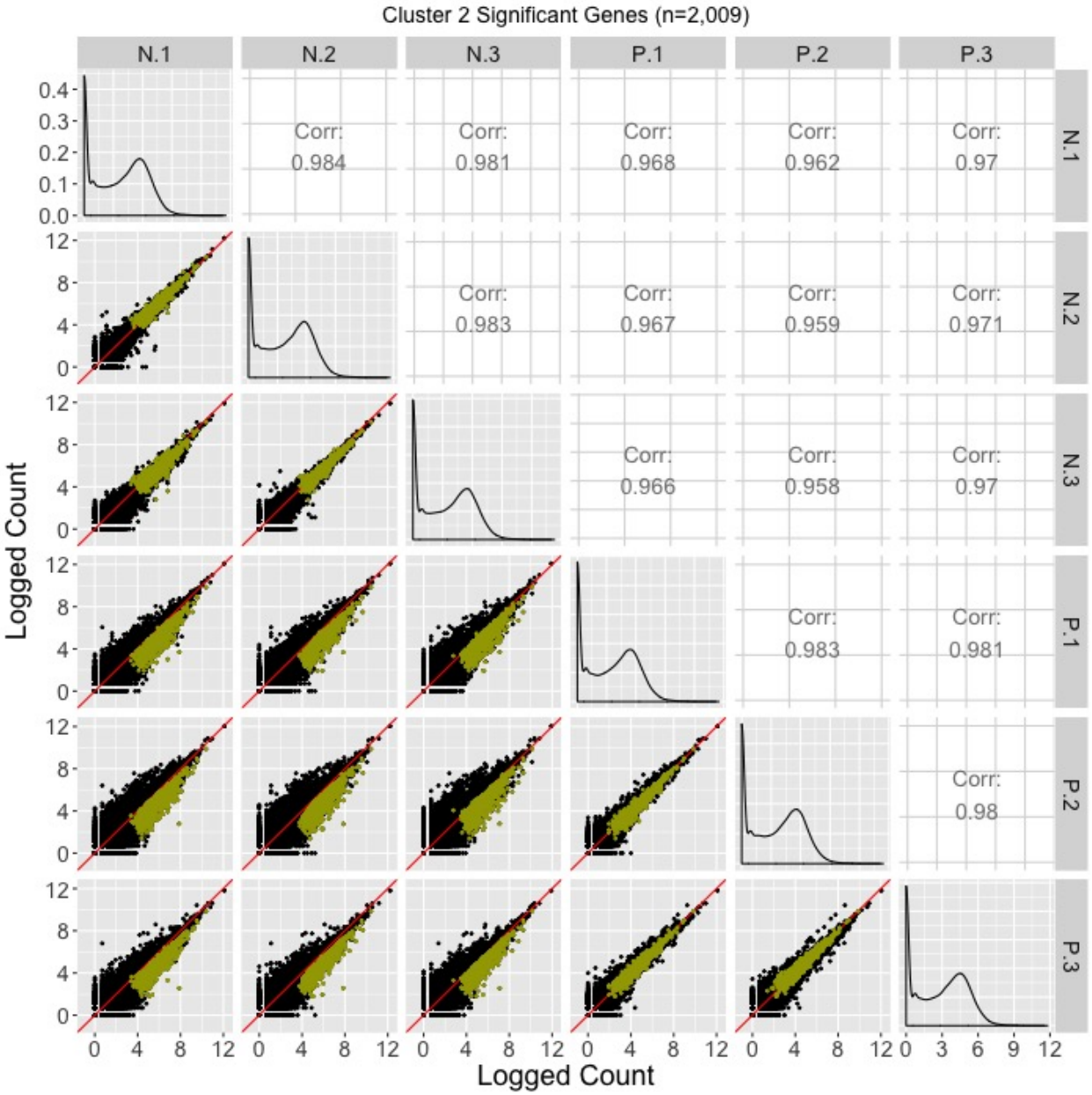

Figure 3: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2009 significant genes in Cluster 2 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in mustard over the scatterplot matrix. They follow the expected patterns of differential expression with most mustard points falling along the x=y line in the scatterplots between replicates, but deviating from the x=y line in the scatterplots between treatments. The deviation consistently demonstrates higher expression in the N group than in the P group. Hence, these mustard points seem to represent genes that were significantly overexpressed in the N group, which draws the same conclusion with what we derived using the parallel coordinate plots in Figure 2 of the paper. One difficulty with plotting such a large number of DEGs onto the scatterplot matrix is that overplotting can obscure our inability to determine how many DEGs are in a given location.

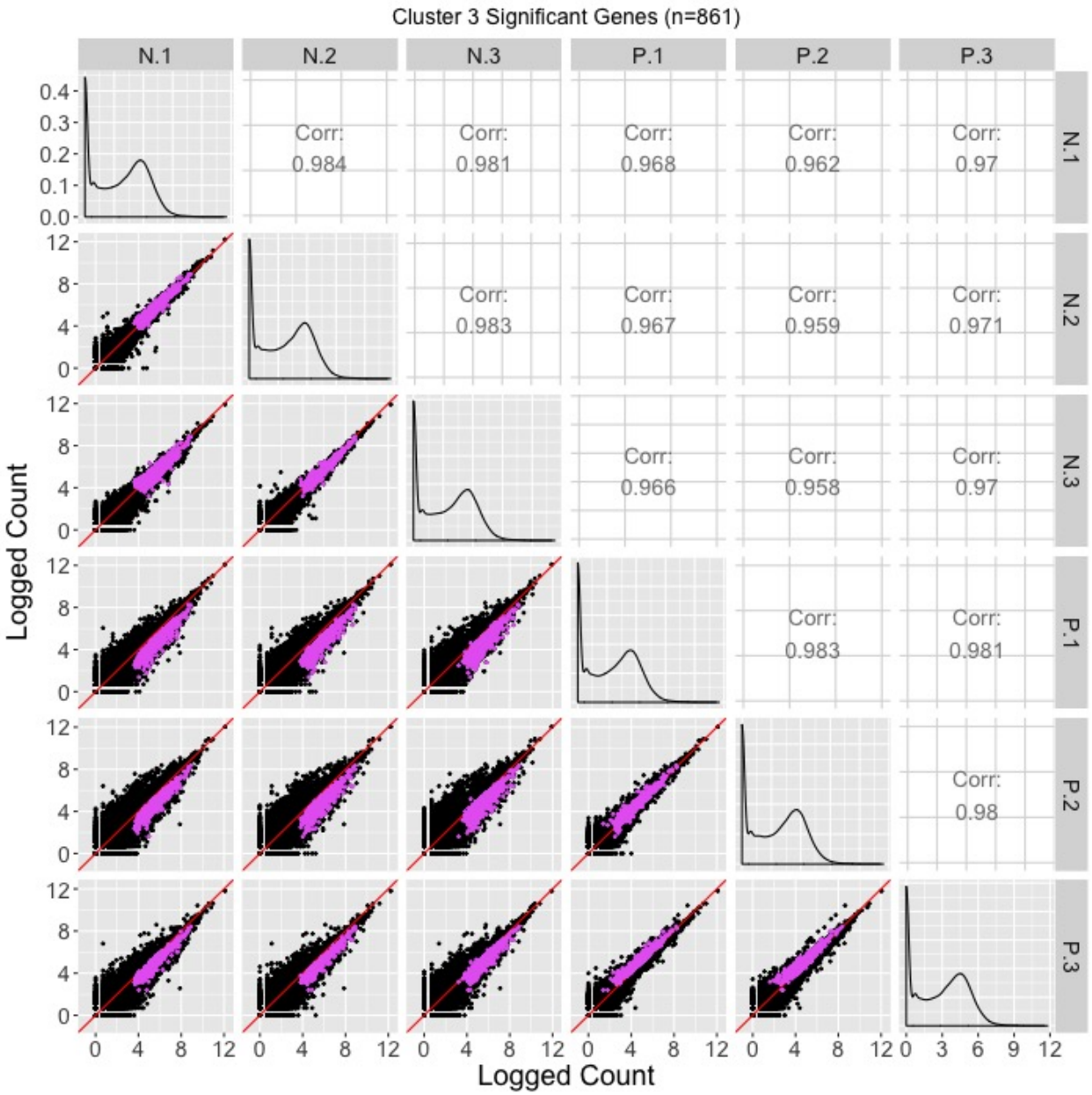

Figure 4: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 861 significant genes in Cluster 3 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in pink over the scatterplot matrix. For the most part, they follow the expected patterns of differential expression with pink points falling along the x=y line in the scatterplots between replicates, but deviating from the x=y line in the scatterplots between treatments. The deviation consistently demonstrates higher expression in the N group than in the P group. However, the scatterplot between replicates P.1 and P.3 show slightly higher expression in P.3, and the scatterplot between replicates P.2 and P.3 also show slightly higher expression in P.3. Hence, these pink points seem to represent genes that were significantly overexpressed in the N group, but with slight inconstencies in the replicates in the P group. The parallel coordinate plots in Figure 2 of the paper showed this same conclusion and perhaps more clearly. One difficulty with plotting such a large number of DEGs onto the scatterplot matrix is that overplotting can obscure our inability to determine how many DEGs are in a given location.

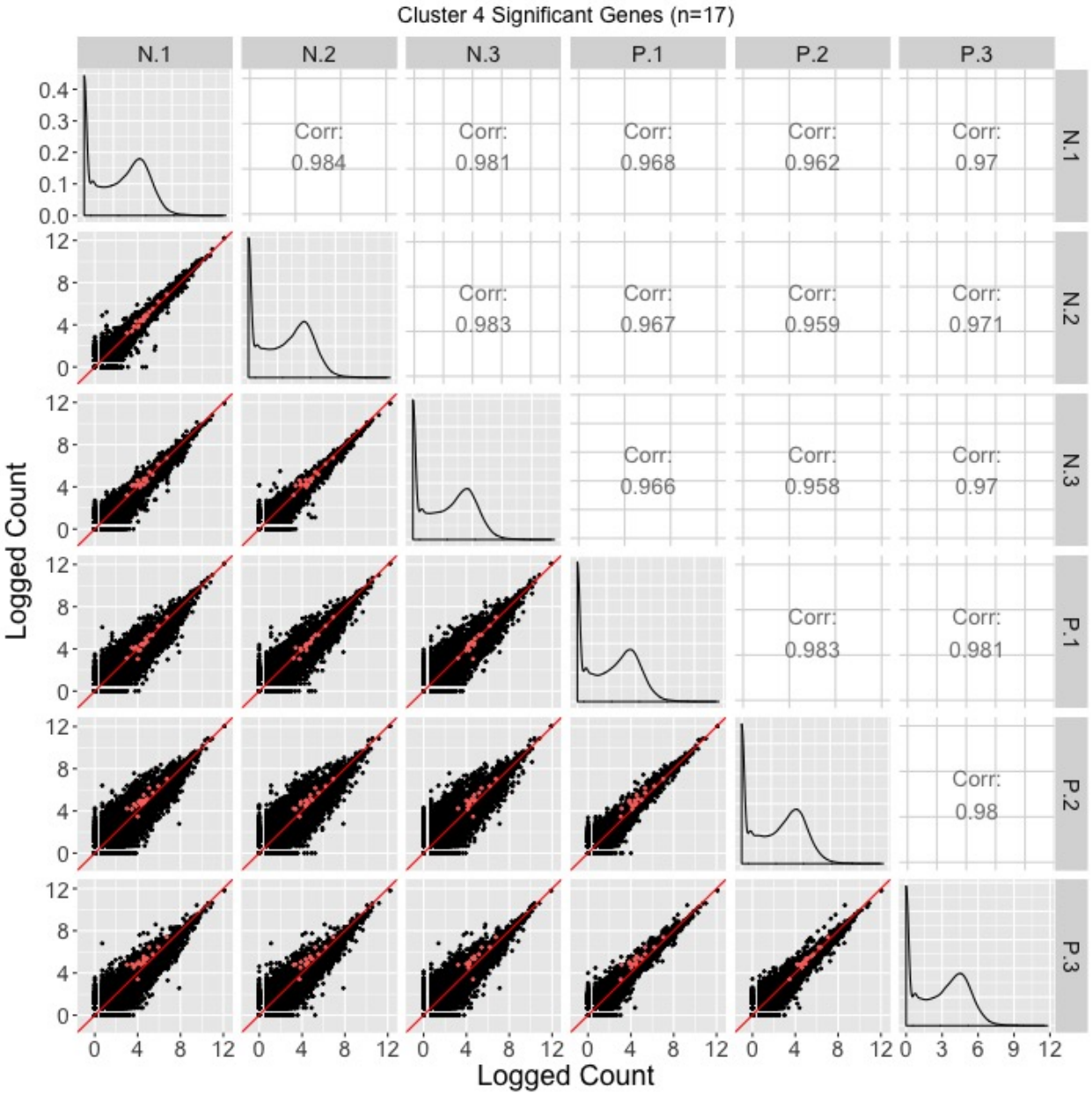

Figure 5: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 17 significant genes in Cluster 4 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in coral over the scatterplot matrix. For the most part, they do not seem to follow the expected patterns of differential expression: In many of the scatterplots between treatments, the coral points do not seem to deviate much from the x=y line. Moreover, in the scatterplots between P.1 and P.2 as well as P.1 and P.3, the coral points seems to indicate an underexpression of the P.1 replicate. We found a similar finding of somewhat messy looking DEG calls in Cluster 4 from Figure 2 in the paper.

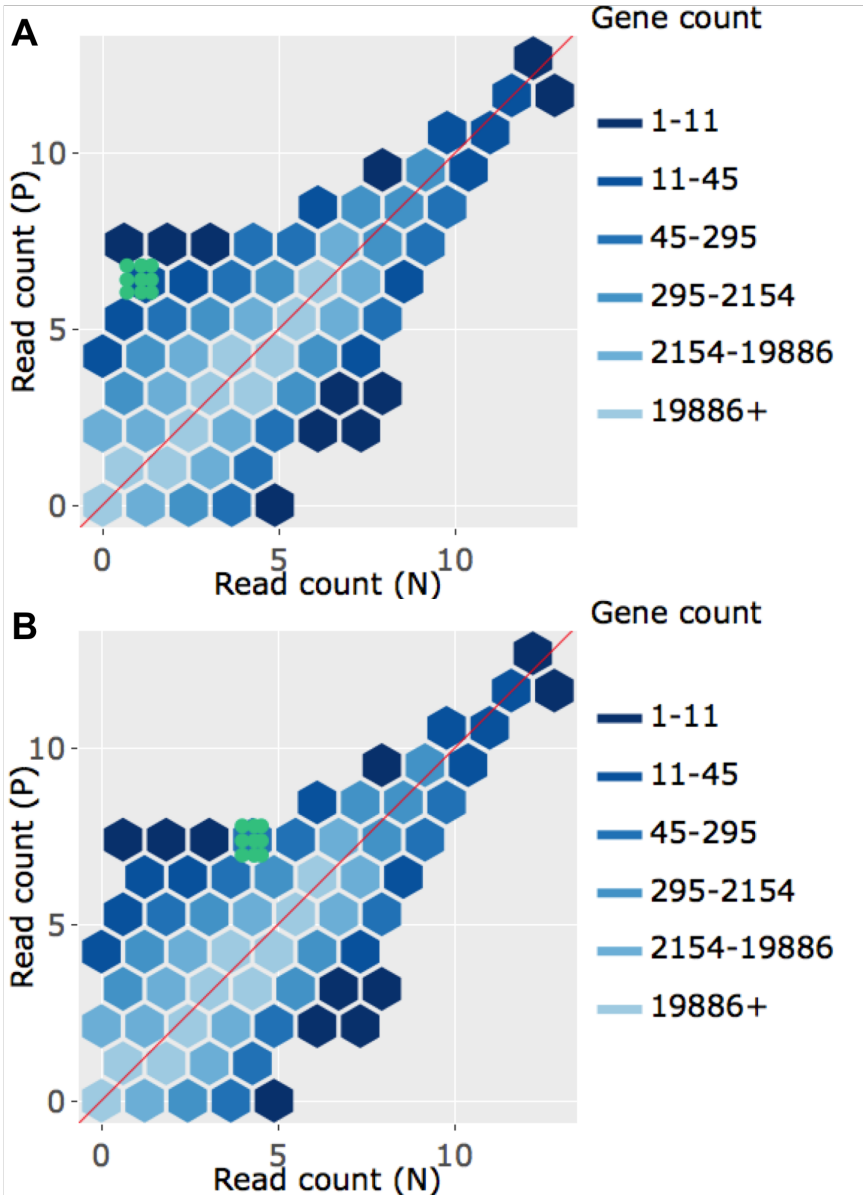

Figure 6: Litre plots for significant genes inside Cluster 1 from Figure 2 of the paper. Subplots A and B each overlay a significant gene from Cluster 1 as nine green points. The genes show a pattern expected of a differentially-expressed one, by clumping together and deviating from the x=y line. Moreover, the genes appear over-expressed in the P group. This is consistent with what we saw in Figure 2 of the paper. To interactively view the litre plot for all significant genes within Cluster 1, please visit https://rnaseqvisualization.shinyapps.io/litreCluster1.

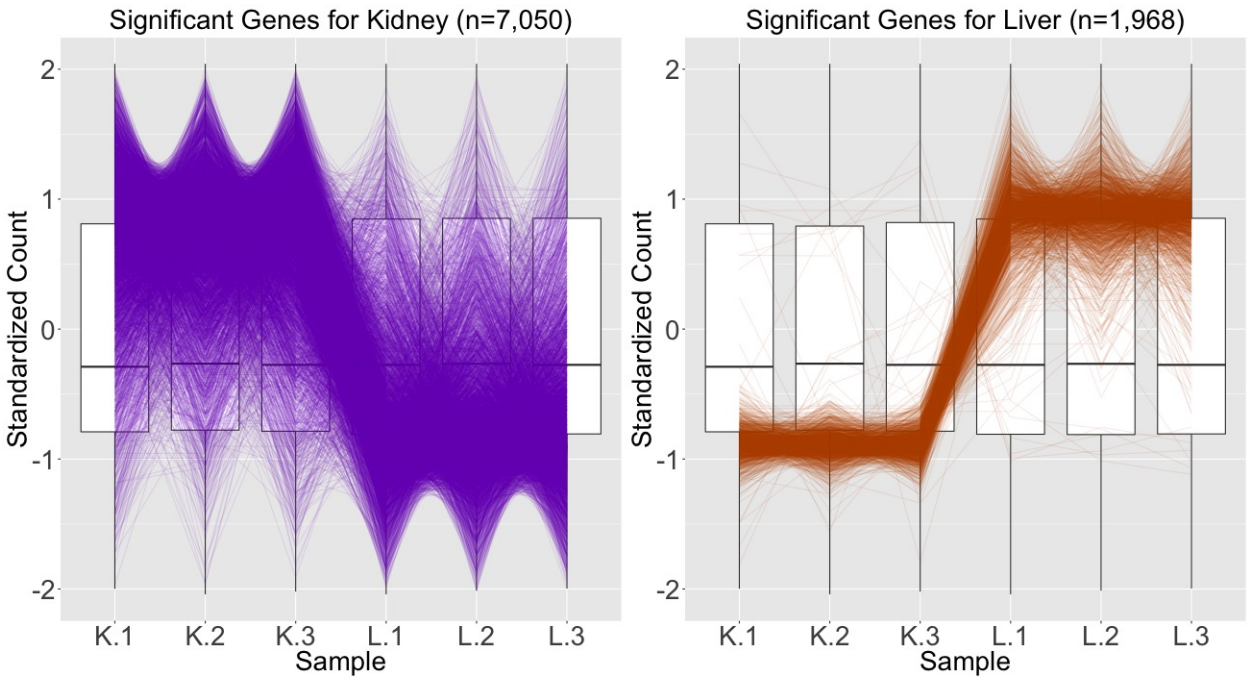

Figure 7: Parallel coordinate plots of the DEGs from liver and kidney technical replicates after standard library scale normalization. The division of DEGs between the two groups was rather disparate, with 78% of the DEGs being kidney-specific and only 22% of the DEGs being liver-specific. Also of note, while the parallel coordinate patterns of the liver-specific DEGs appear as expected, the patterns of the kidney-specific DEGs seem to show comparatively large variability between the replicates.

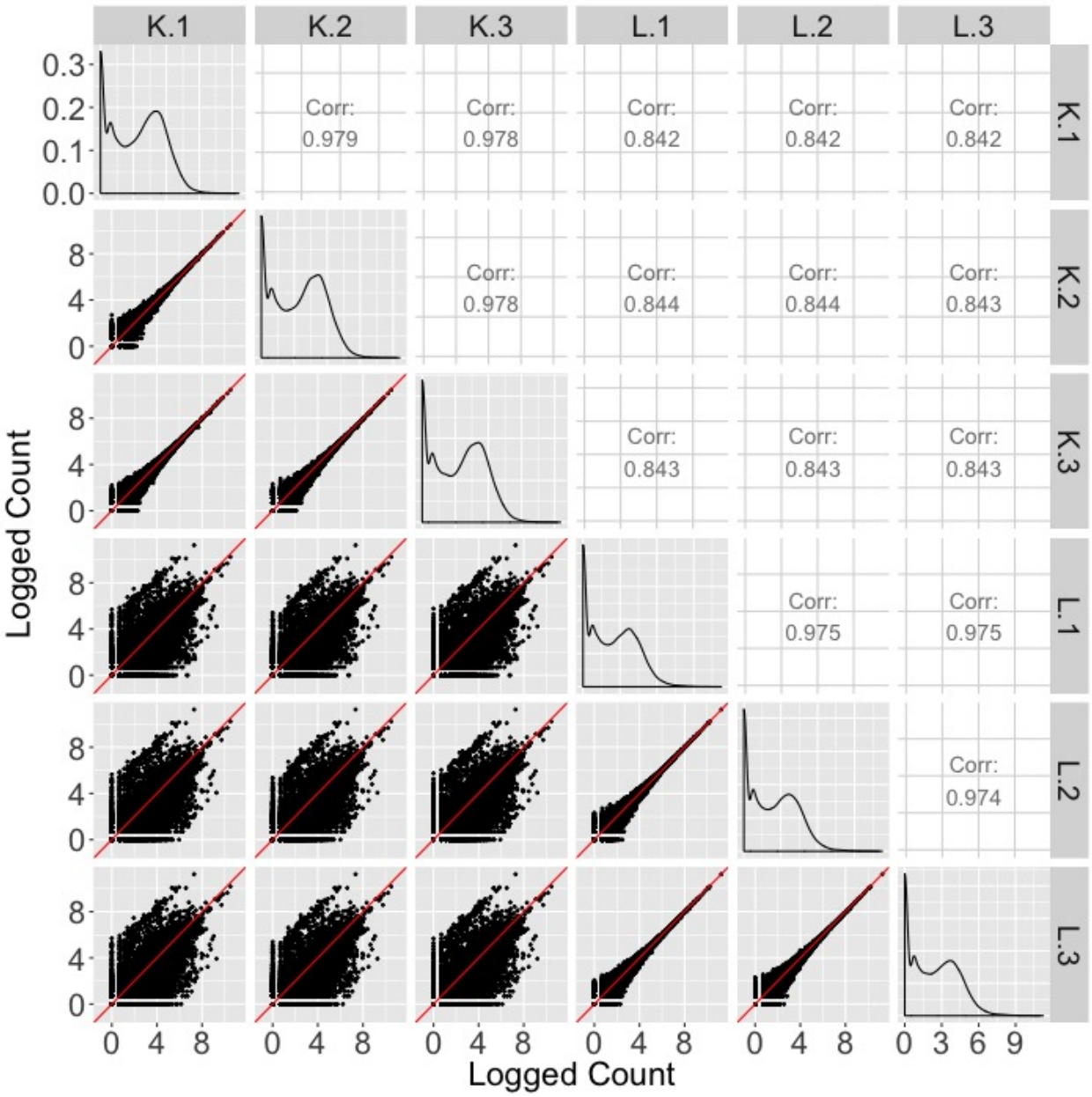

Figure 8: Scatterplot matrix of liver and kidney technical replicates. The technical replicate scatterplots look as precise as expected, with little variability around the x=y line. The treatment group scatterplots have much more variability around the x=y line, as we would expect. However, they each contain a pronounced streak of highly-expressed liver-specific genes, which deviates from the expected distribution.

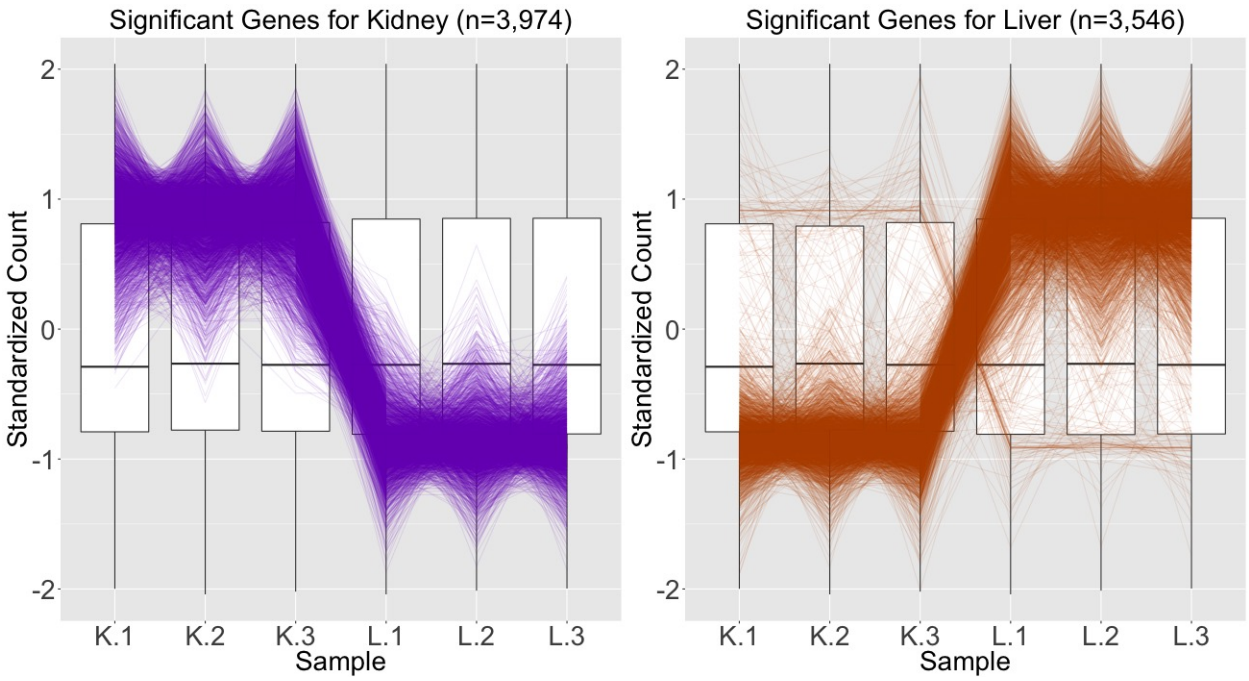

Figure 9: Parallel coordinate plots of the DEGs from liver and kidney technical replicates after TMM normalization. The division of DEGs between the two groups is more balanced than in Figure 7, with 53% of the DEGs being kidney-specific and 47% of the DEGs being liver-specific. Additionally, the parallel coordinate patterns of both the liver-specific and kidney-specific DEGs appear more consistent and as expected.

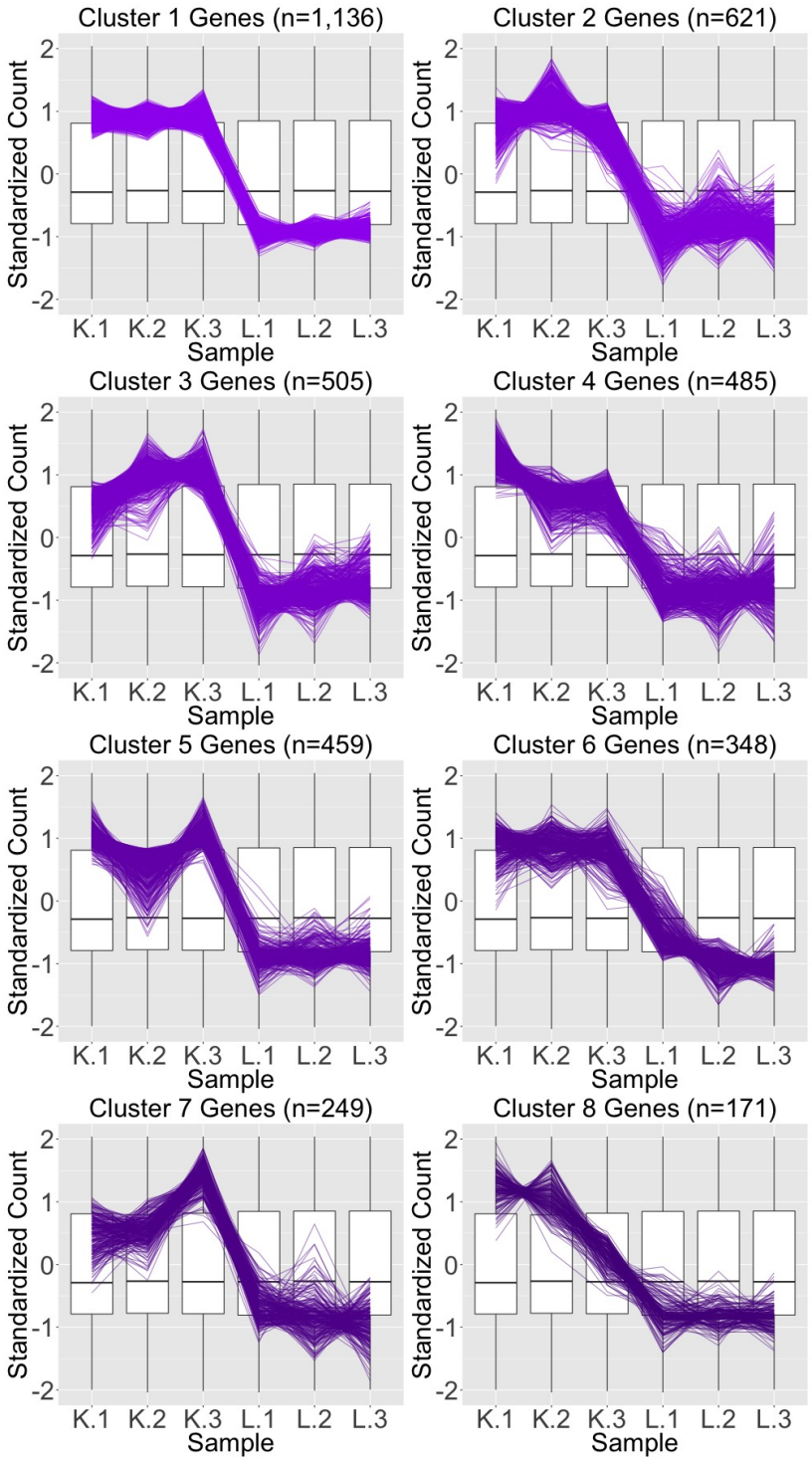

Figure 10: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 3,974 genes that remained in the kidney-specific DEGs after TMM normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. The ideal pattern of DEGs is especially captured in the first cluster (the largest one with 1,136 genes).

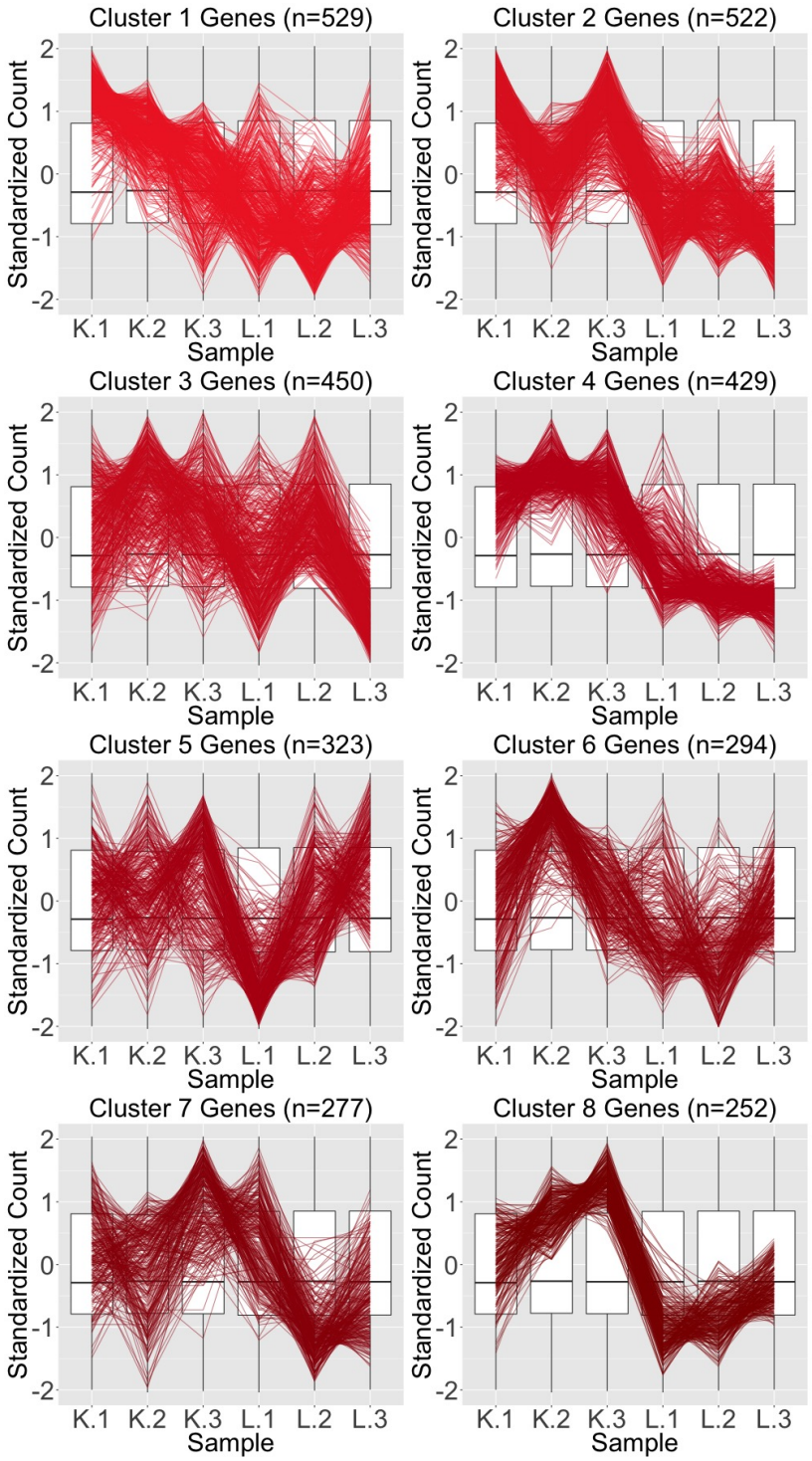

Figure 11: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 3,076 genes that were removed from the kidney-specific DEGs after TMM normalization. Unlike in Figure 10, the patterns in almost all clusters do not resemble the expected DEG format; instead, they show large variability between replicates and small variability between groups. In some clusters, it is difficult to even determine which treatment group would be the overexpressed one if its genes were in fact DEGs. Taken together, this plot provides additional statistical evidence that the application of TMM normalization successfully removed genes that were previously mislabeled as DEGs (in Figure 7) with library scaling normalization.

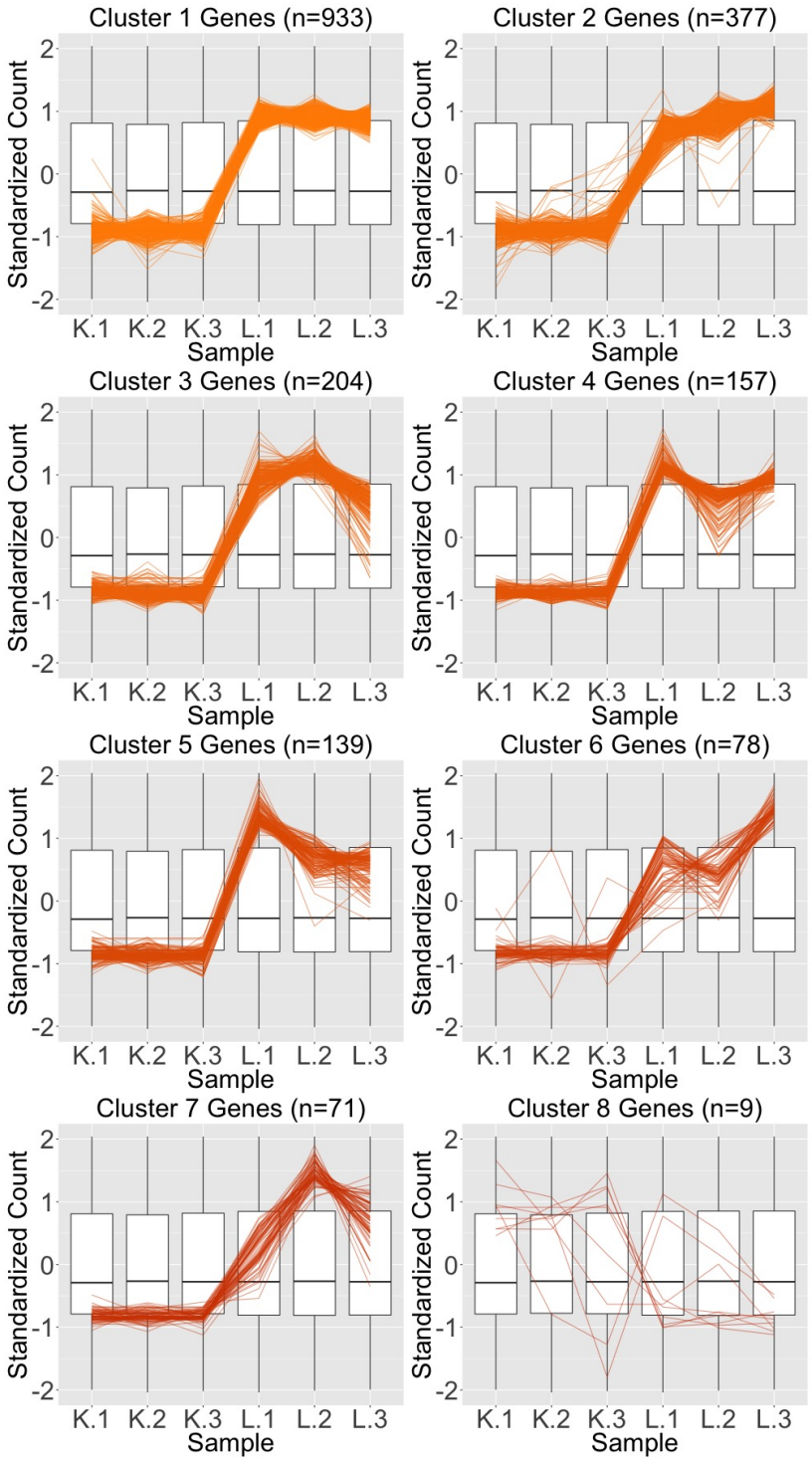

Figure 12: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 1,968 genes that were designated DEGs after library scale normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. The ideal pattern of DEGs is especially captured in the first cluster (the largest one with 933 genes).

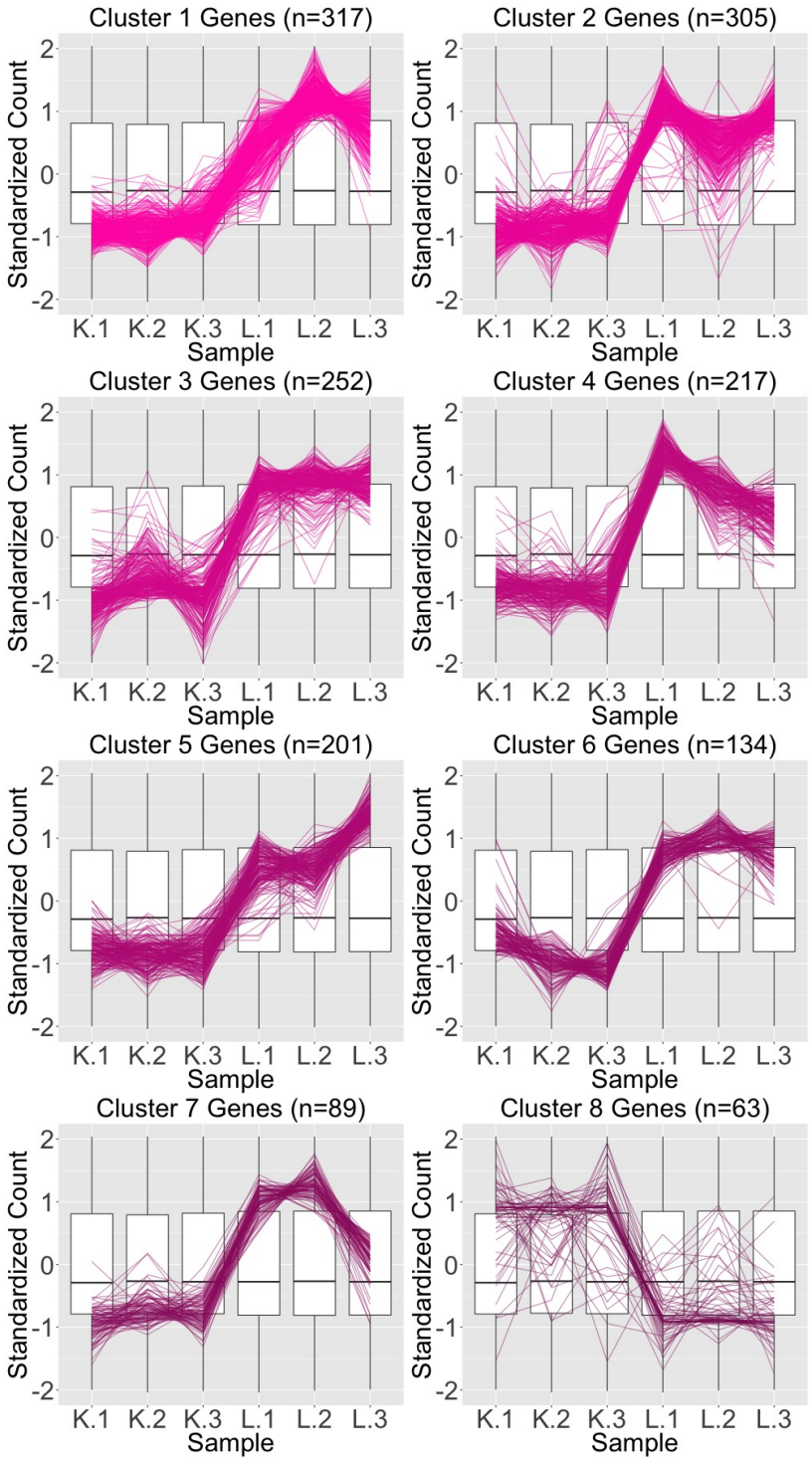

Figure 13: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 1,968 genes that were designated DEGs after library scale normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. The ideal pattern of DEGs is especially captured in the first cluster (the largest one with 933 genes).