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

Lindsay Rutter

March 16, 2018

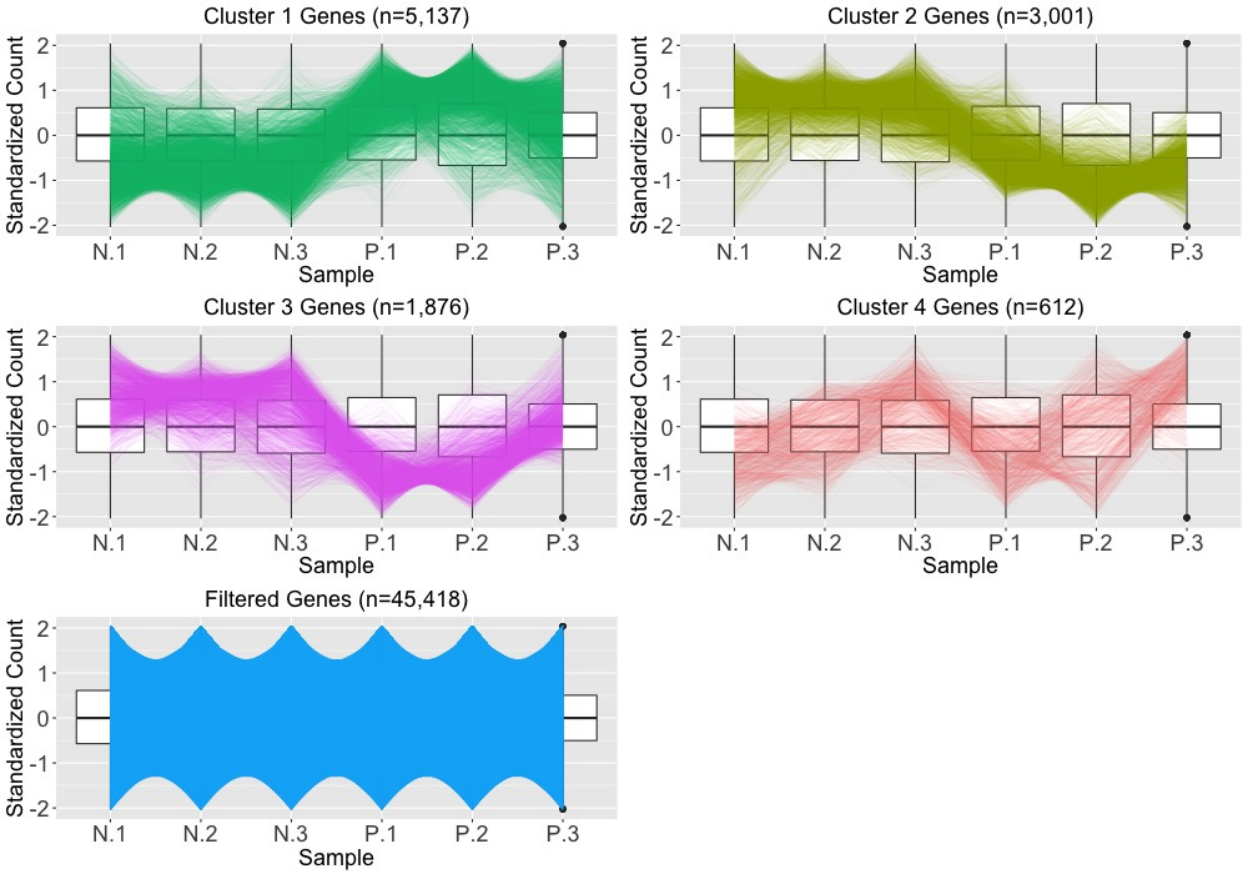

Figure 1: Example application of parallel coordinate plots using the iron-metabolism soybean dataset. We filtered genes with low means and/or variance, performed a hierarchical clustering analysis with a cluster size of four, and visualized the results using parallel coordinate lines. Most non-filtered 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.

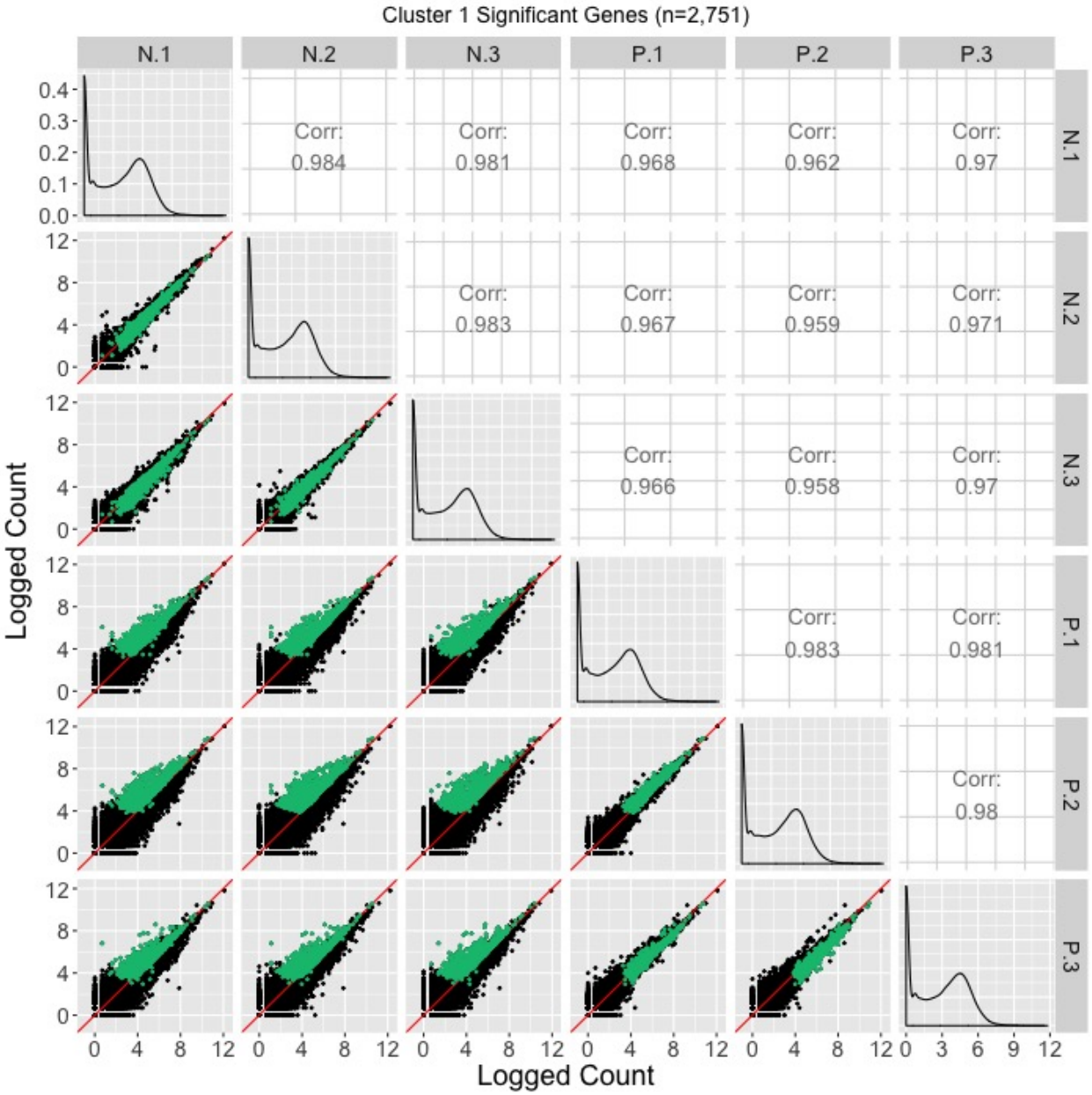

Figure 2: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2751 significant genes in Cluster 1 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in green over the 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.

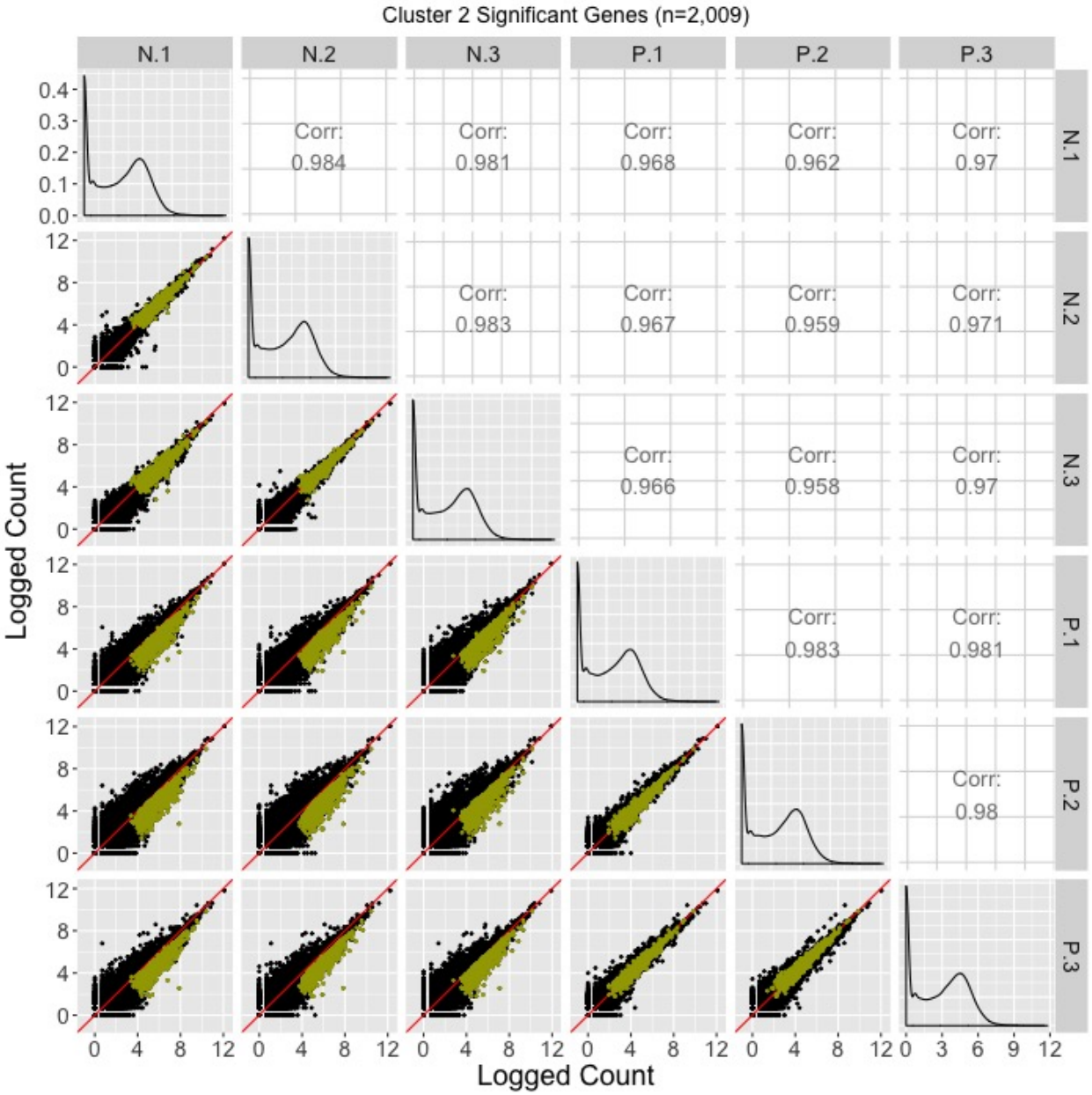

Figure 3: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2009 significant genes in Cluster 2 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in mustard over 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.

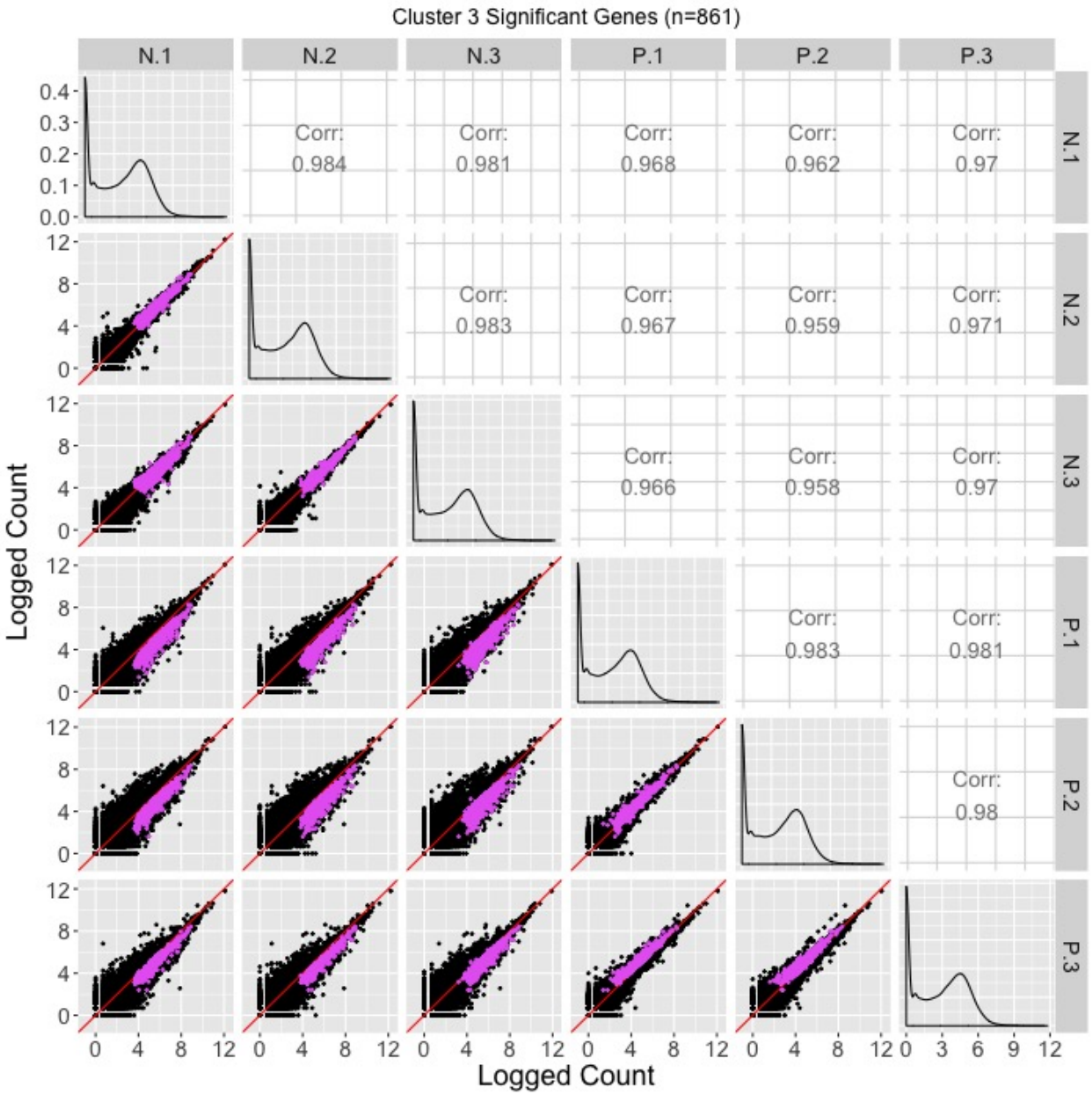

Figure 4: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 861 significant genes in Cluster 3 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in pink over the 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.

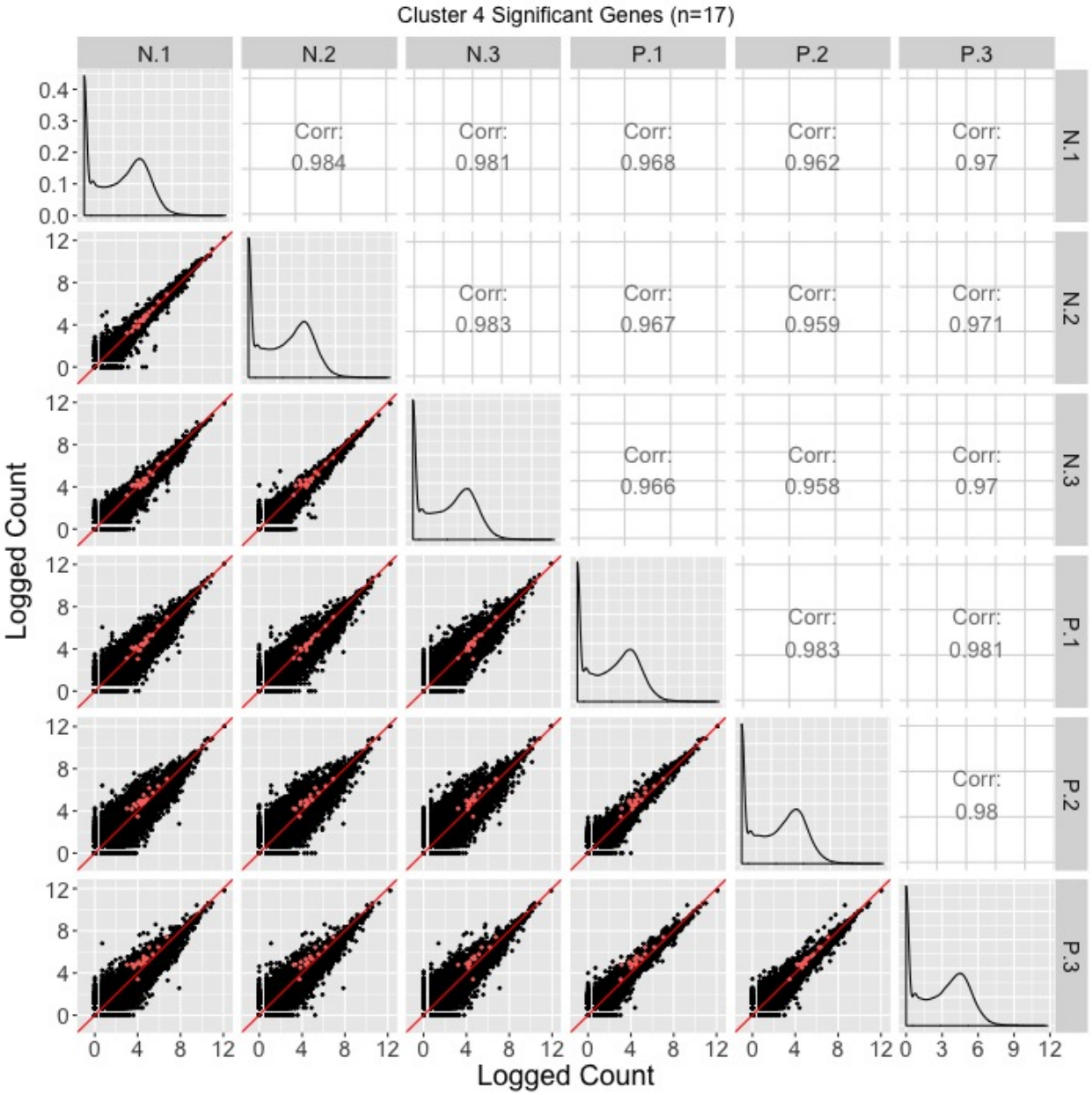

Figure 5: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 17 significant genes in Cluster 4 after performing a hierarchical clustering analysis with a cluster size of four. These significant genes are overlaid in coral over the 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.

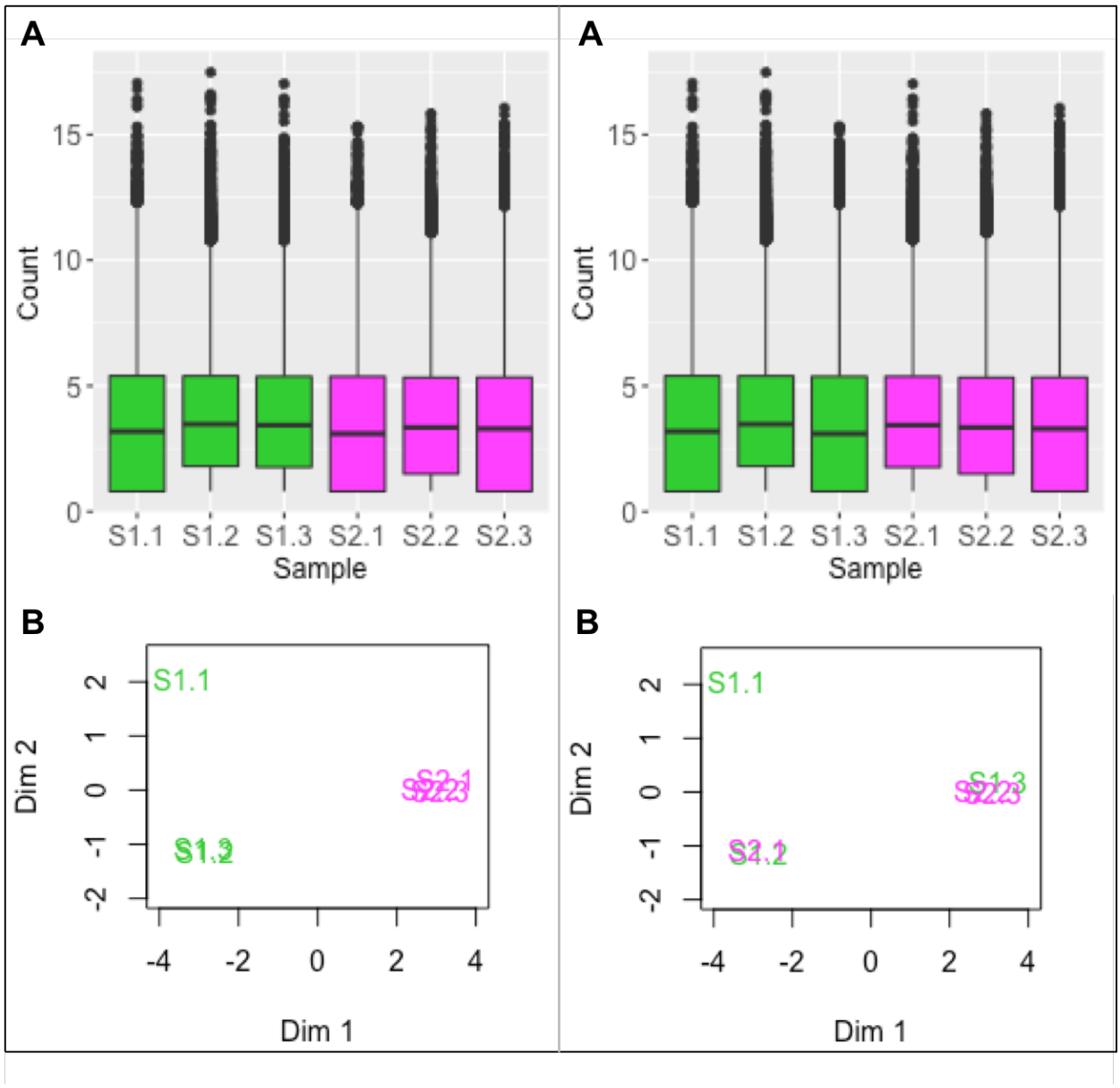

Figure 6: Boxplots and MDS plots are popular plotting tools for RNA-seq analysis. This figure shows these traditional plots applied to the cotyledon data before sample switching (left half) and after sample switching (right half). We cannot suspect from the right boxplot that samples S1.3 and S2.1 have been swapped (subplots A). This is because all six samples have similar five number summaries. For the MDS plots, we do see a cleaner separation of the two treatment groups across the first dimension in the left plot than in the right plot (subplots B). However, taking into account the second dimension, both MDS plots contain three clusters, with sample S1.1 appearing in its own cluster. Without seeing one distinct cluster for each of the two treatment groups, it is difficult to suspect that samples S1.3 and S2.1 have been swapped in the right MDS plot (subplots B). We can only derive clear suspicion that the samples may have been switched by using plots that provide gene-level resolution like with the scatterplot matrix (Figure **??**).

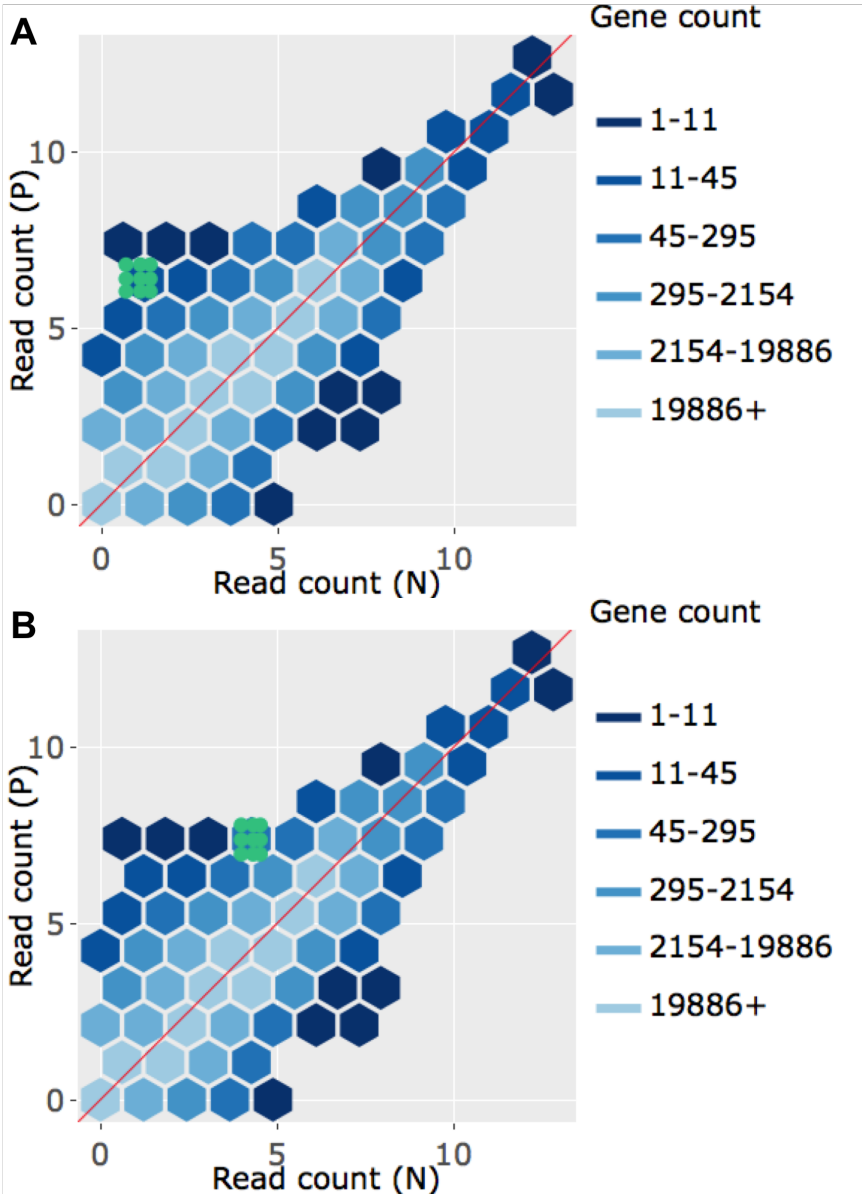

Figure 7: Litre plots for significant genes inside Cluster 1 from Figure 2 of the paper. Subplots A and B each overlay a significant gene from Cluster 1 as nine green points. The genes show a pattern expected of a differentially-expressed one, by clumping together and deviating from the x=y line. Moreover, 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.

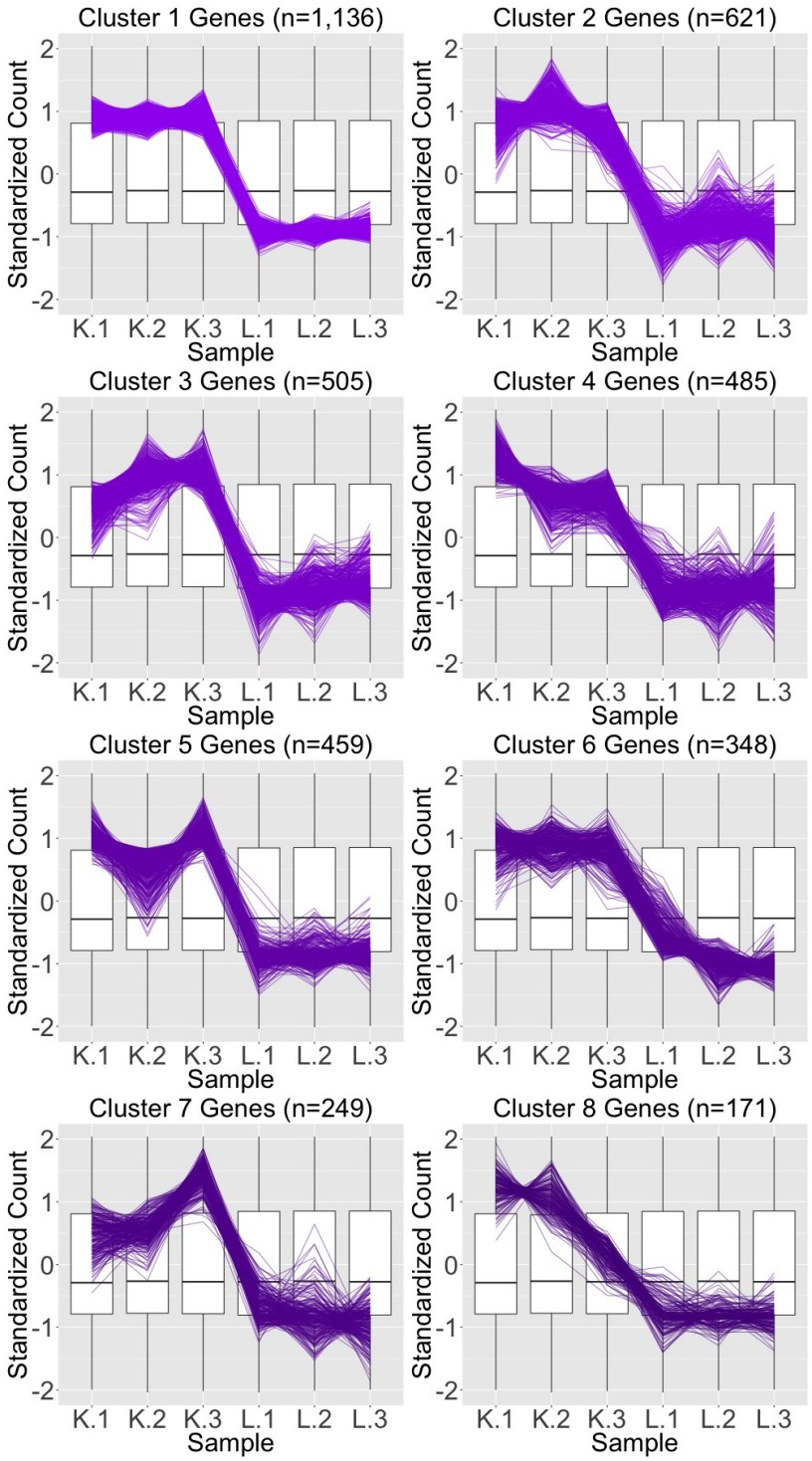

Figure 8: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 3,974 genes that remained in the kidney-specific DEGs after TMM normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. The ideal pattern of DEGs is especially captured in the first cluster (the largest one with 1,136 genes). We used hierarchical clustering to mitigate additional overplotting that would occur if we were to plot all genes onto only one parallel coordinate plot.

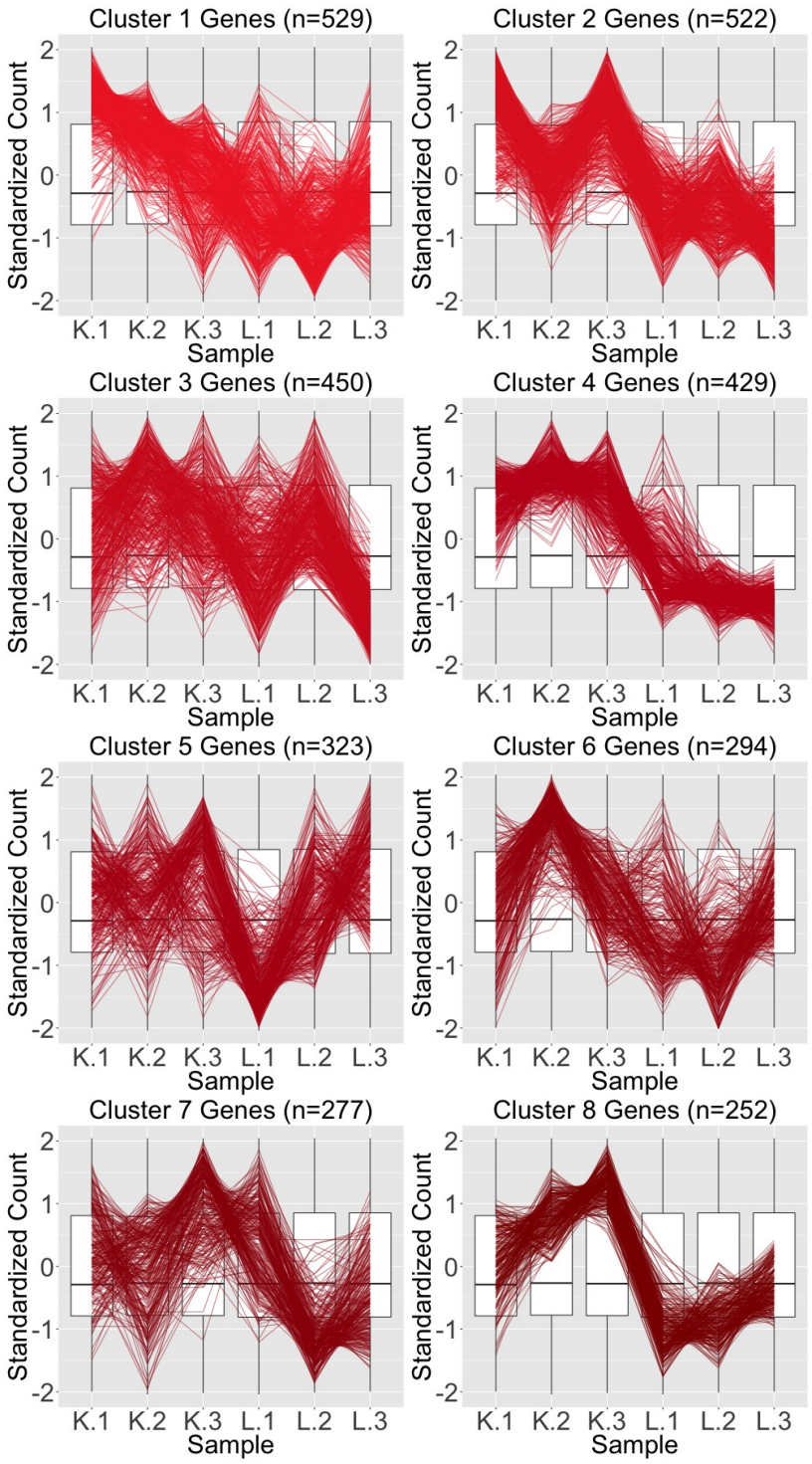

Figure 9: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 3,076 genes that were removed from the kidney-specific DEGs after TMM normalization. Unlike in Figure **??**, the patterns in almost all clusters do not resemble the expected DEG format; instead, they show large variability between replicates and small variability between groups. In some clusters, it is difficult to even determine which 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 kidney-specific DEGs (in Figure **??**) with library scaling normalization. We used hierarchical clustering to mitigate additional overplotting that would occur if we were to plot all genes onto only one parallel coordinate plot.

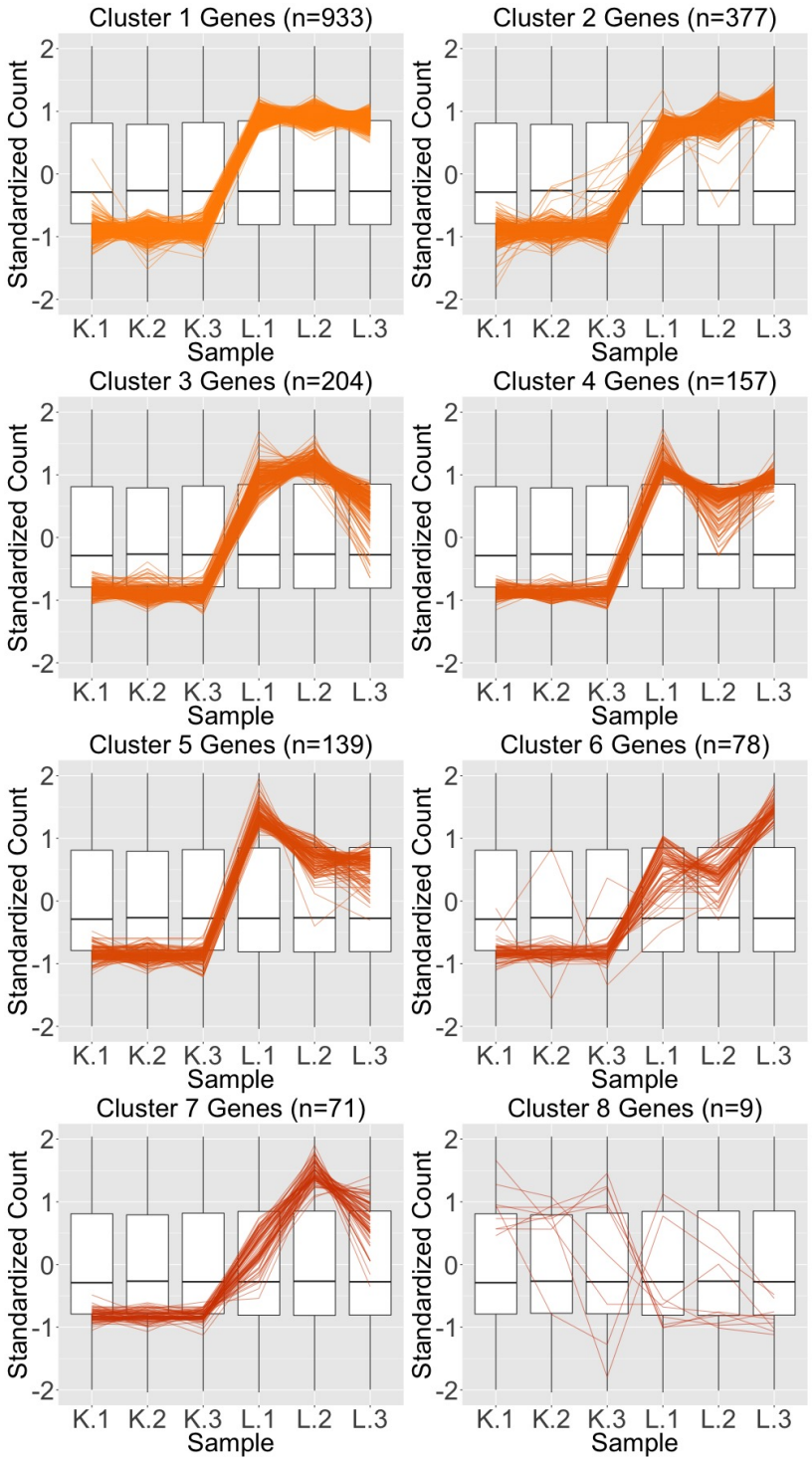

Figure 10: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 1,968 genes that were initially designated liver-specific DEGs after library scale normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. The ideal pattern of DEGs is especially captured in the first cluster (the largest one with 933 genes). We used hierarchical clustering to mitigate additional overplotting that would occur if we were to plot all genes onto only one parallel coordinate plot.

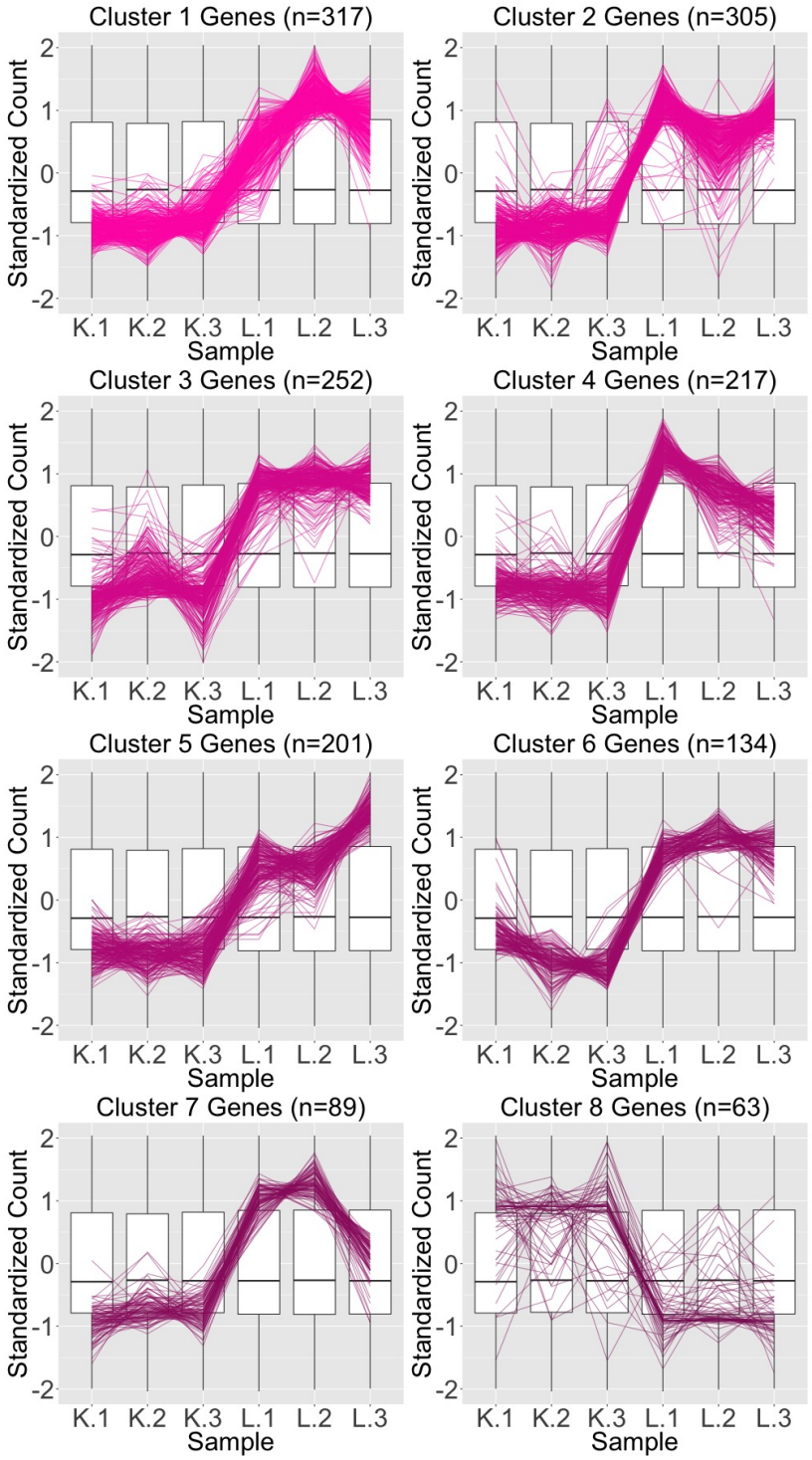

Figure 11: Parallel coordinate plots showing hiearchical clustering analysis results of size eight for the 1,578 genes that were *added* as liver-specific DEGs after TMM normalization. We see that, for the most part, the parallel coordinate patterns follow the expected patterns across the clusters. We used hierarchical clustering to mitigate additional overplotting that would occur if we were to plot all genes onto only one parallel coordinate plot.

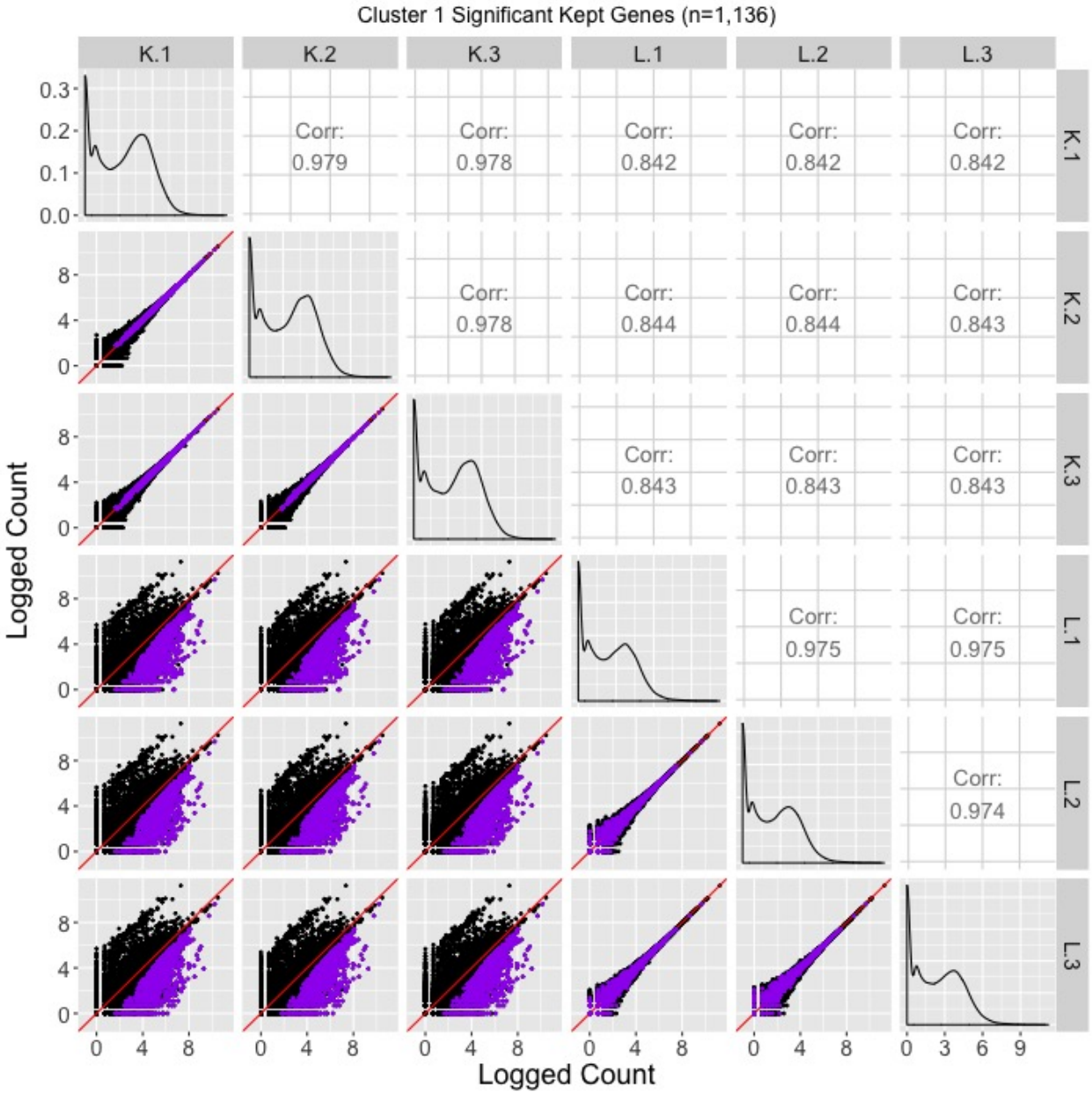

Figure 12: Scatterplot matrix of the 1,136 genes that were in the first cluster (of Figure **??**) from genes that remained as kidney-specific DEGs even after TMM normalization. With this scatterplot matrix, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

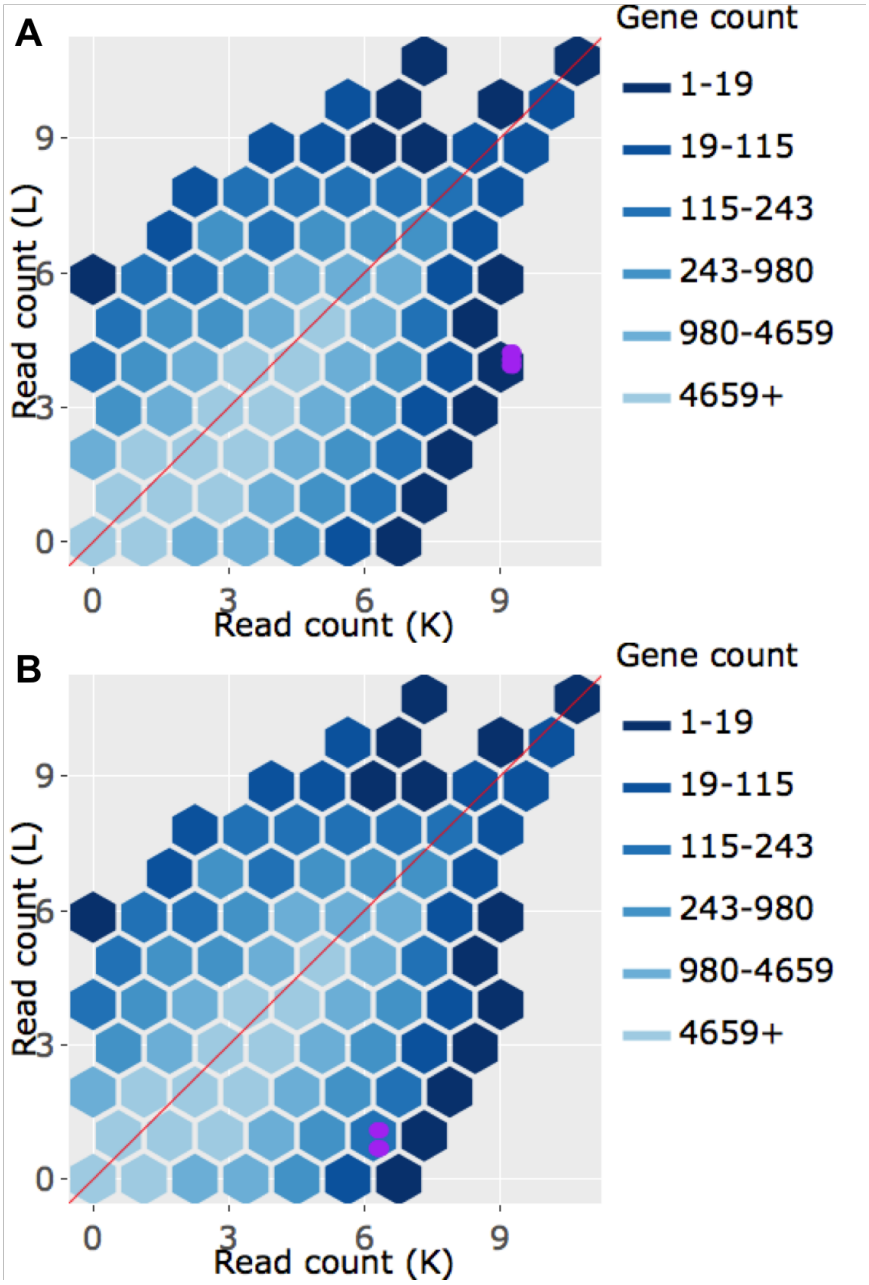

Figure 13: Example litre plots from the 1,136 genes that were in the first cluster (of Figure **??**) of genes that remained kidney-specific DEGs even after TMM normalization. With these litre plots, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

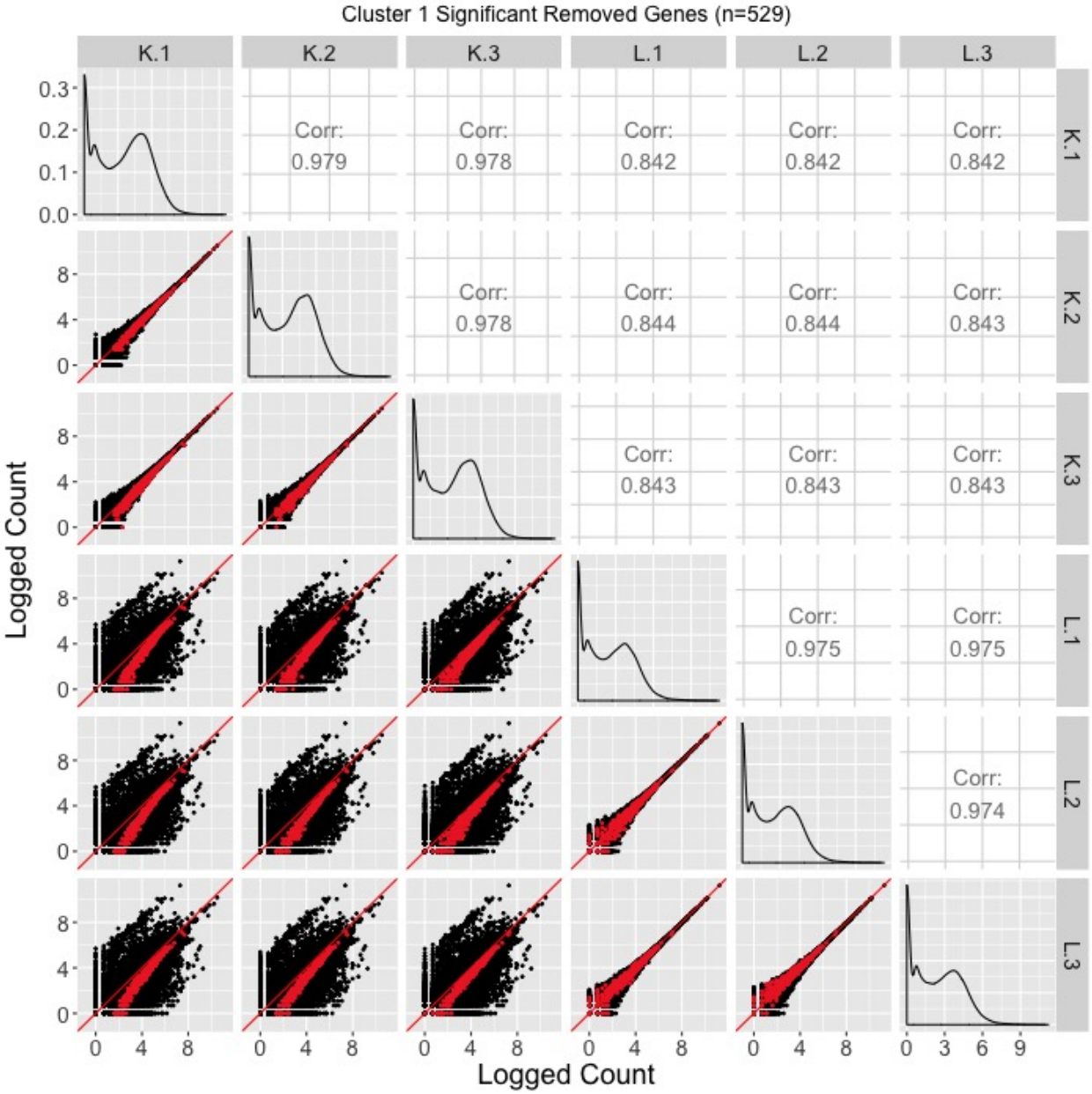

Figure 14: Scatterplot matrix of the 529 genes that were in the first cluster (of Figure **??**) from genes that no longer remained as kidney-specific DEGs after TMM normalization. With this scatterplot matrix, we verify from an additional perspective that these genes do not demonstrate the expected patterns of DEGs too strongly (they do not deviate much from the x=y line in the treatment scatterplots). This provides additional evidence that TMM normalization removing these genes from DEG status may be valid.

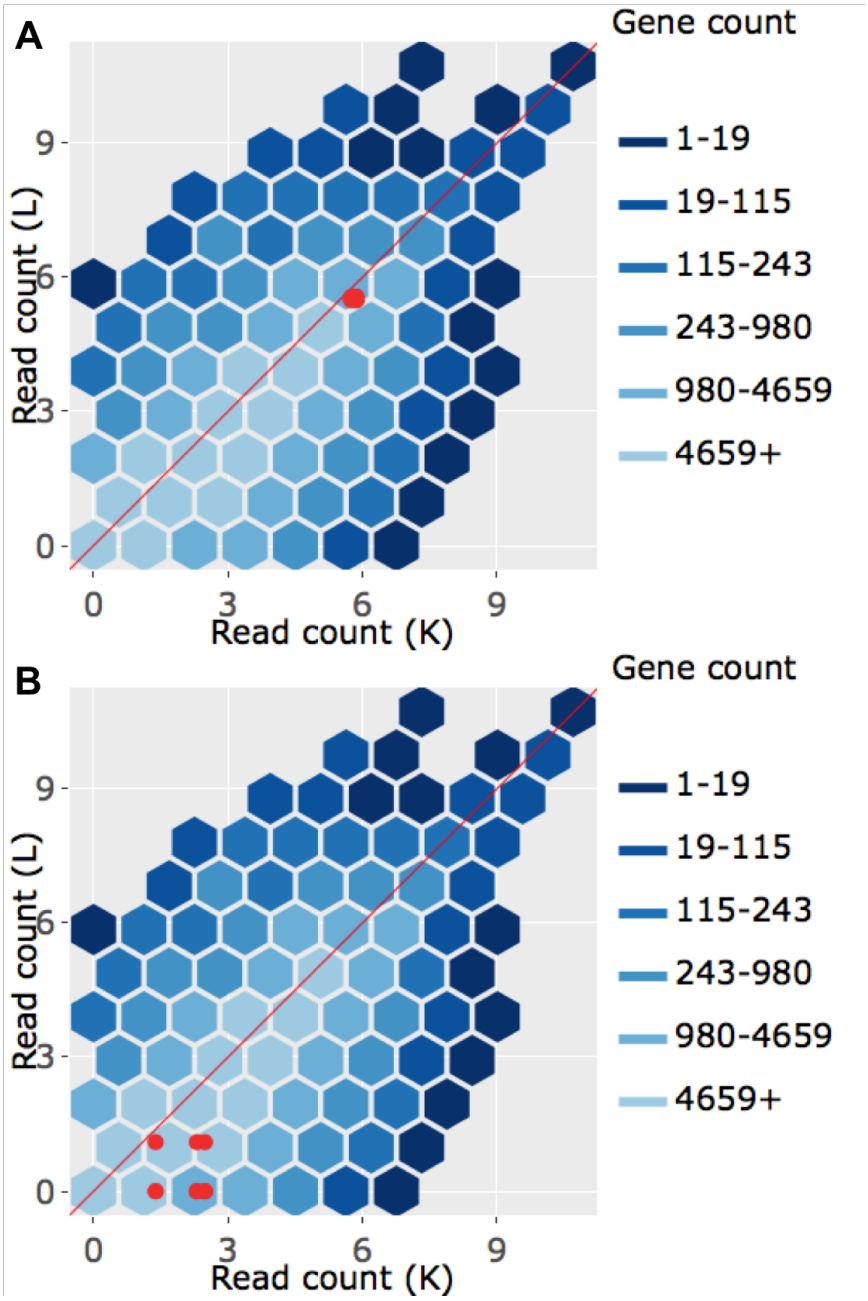

Figure 15: Example litre plots from the 529 genes that were in the first cluster (of Figure **??**) of genes that no longer remained as kidney-specific DEGs after TMM normalization. With these litre plots, we verify from an additional perspective that these genes do not demonstrate the expected patterns of DEGs. This provides additional evidence that TMM normalization removing these genes from DEG status may be valid.

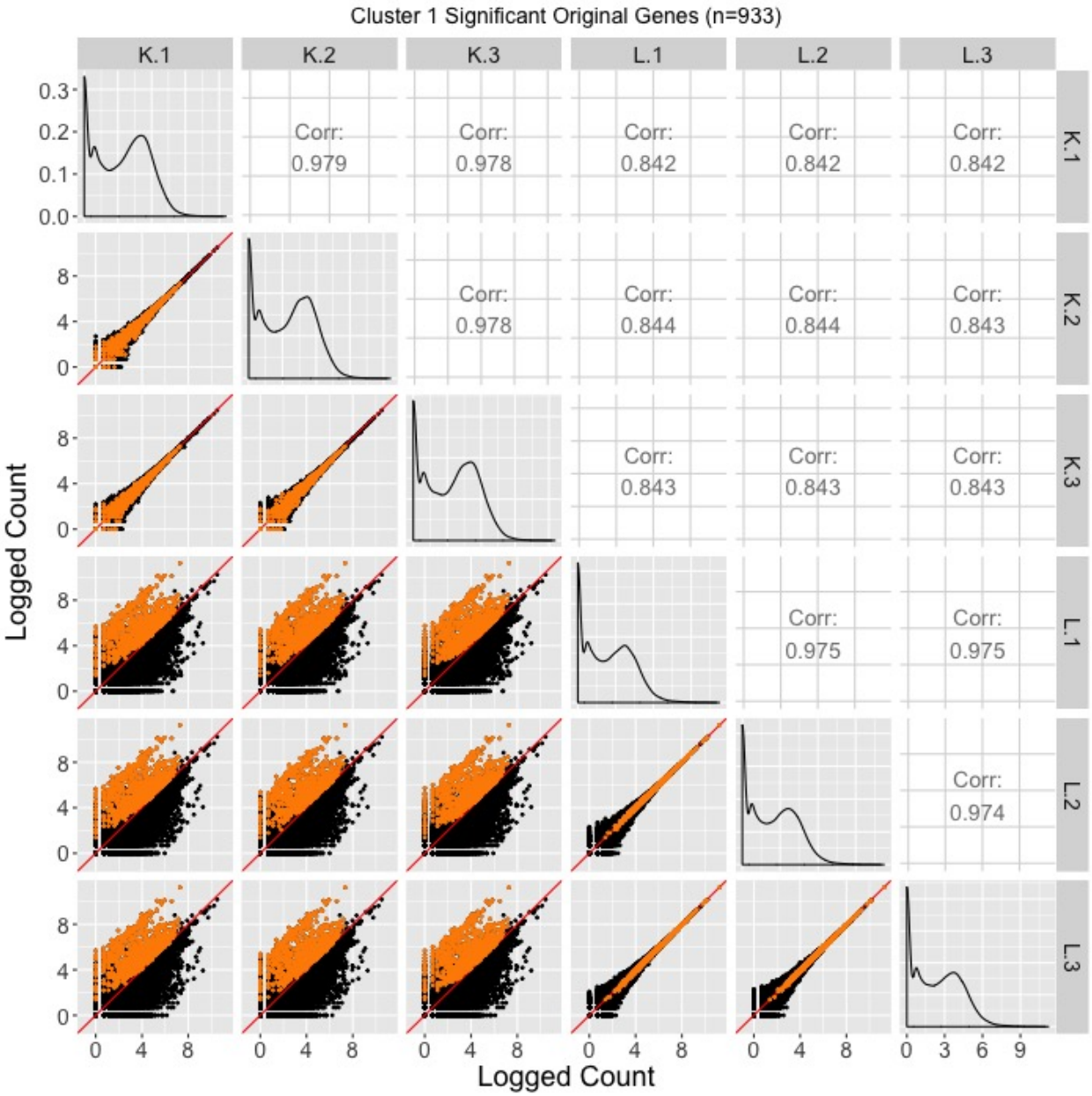

Figure 16: Scatterplot matrix of the 933 genes that were in the first cluster (of Figure **??**) from genes that were initially designated as liver-specific DEGs after library scale normalization. With this scatterplot matrix, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

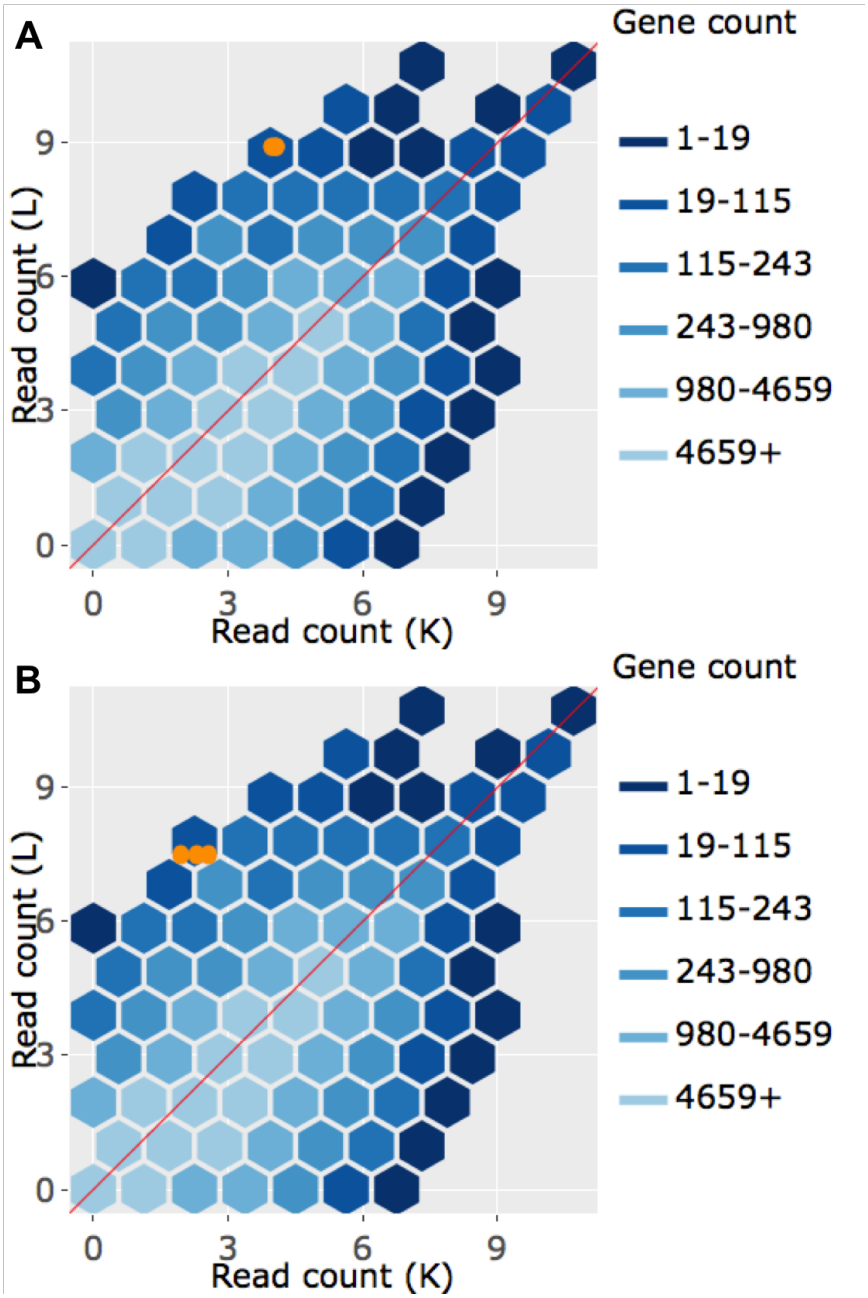

Figure 17: Example litre plots from the 933 genes that were in the first cluster (of Figure **??**) from genes that were initially designated as liver-specific DEGs after library scale normalization. With these litre plots, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

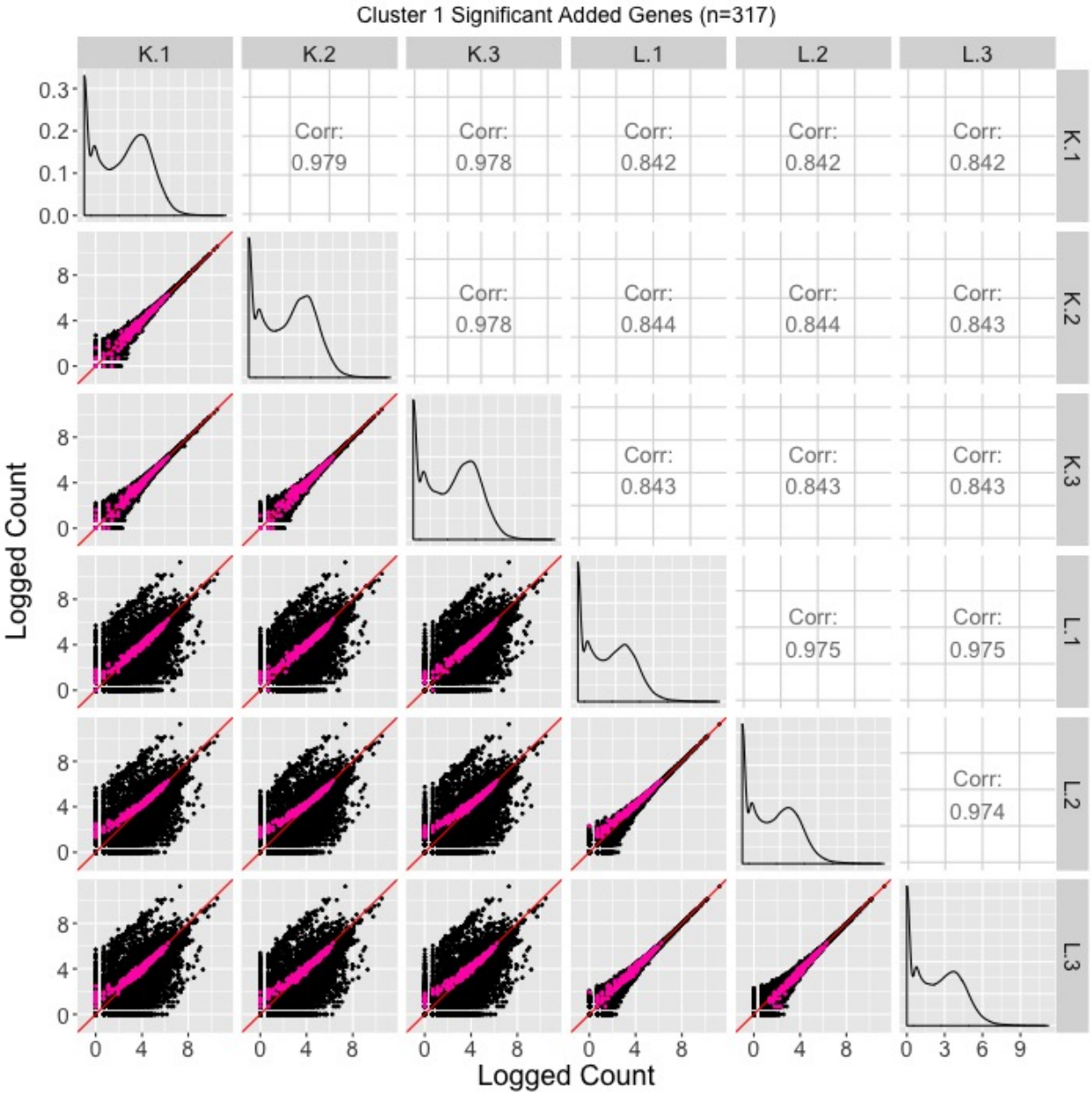

Figure 18: Scatterplot matrix of the 317 genes that were in the first cluster (of Figure **??**) from genes that were *added* as liver-specific DEGs after TMM normalization. With this scatterplot matrix, we are **unable** to verify from an additional perspective that these genes demonstrate the expected patterns of DEGs. (Need to find an explanation for this).

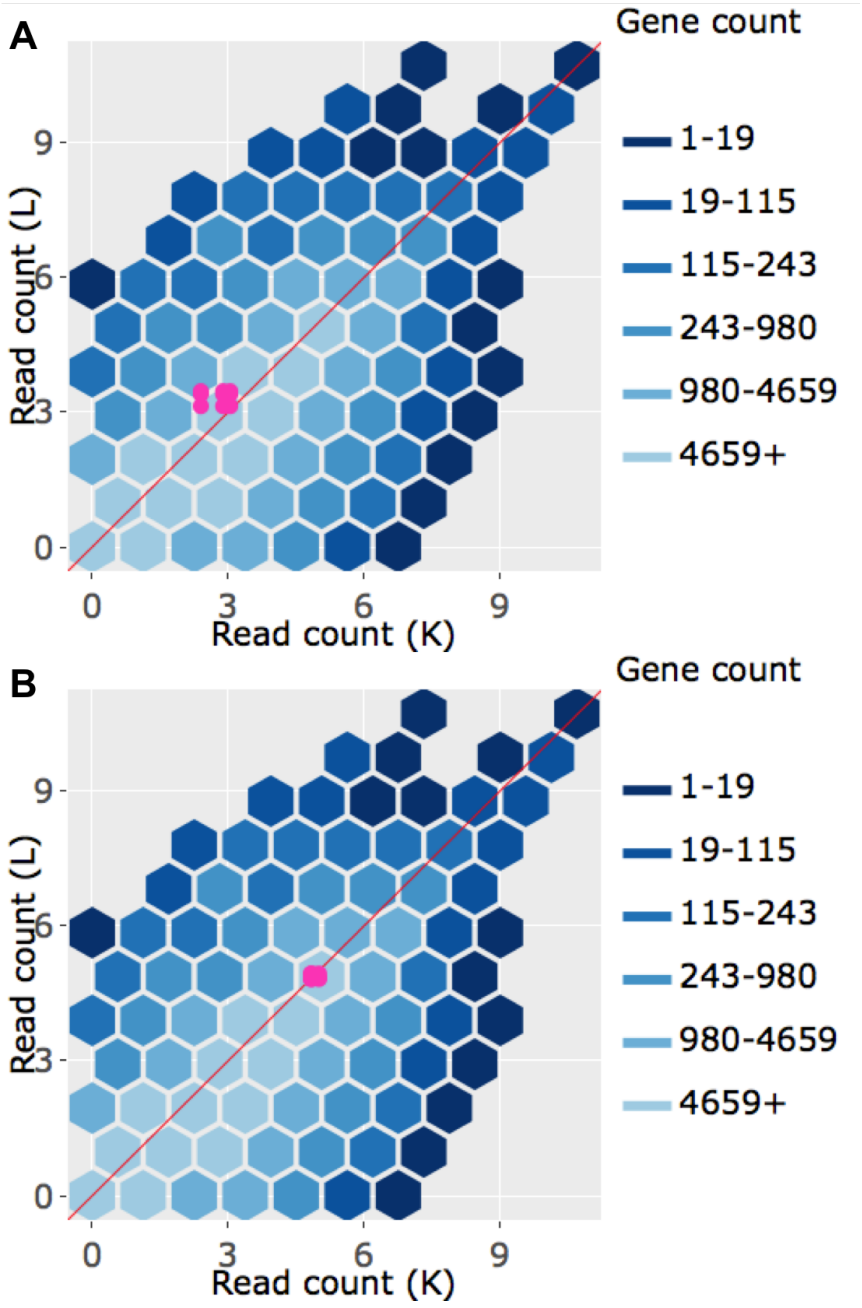

Figure 19: Example litre plots from the 1,578 genes that were in the first cluster (of Figure **??**) from genes that were *added* as liver-specific DEGs after TMM normalization. With these litre plots, we are **unable** to verify from an additional perspective that these genes demonstrate the expected patterns of DEGs. (Need to find an explanation for this).