Appendix: Visualization Methods for RNA-sequencing Data Analysis

Should the reader be interested, this appendix contains supplementary analyses that extend upon the main concepts discussed in the paper. On this page, we provide a brief description of the figures within this appendix:

- Supplementary figure 1 shows that some of the most popular RNA-seq visualization tools cannot always detect common errors, such as sample switching. Side-by-side boxplots and MDS plots are shown for the soybean cotyledon dataset after deliberately switching samples between treatment groups (Brown and Hudson, 2015). For the most part, they are unable to detect the error. In contrast, the less-popular method of scatterplot matrices can quickly detect the possibility of the error (see Figure 5 from the main paper).

- Supplementary figure 2 shows that the number of DEGs drastically changes across three time points when comparing iron sufficient versus iron deficient conditions in soybean leaves (Moran Lauter and Graham, 2016). As a result, the authors of this study postulate that the streak of genes observed in the scatterplot matrix containing the subset of data at the 120 minute time point (Figure 6 in the main paper) may be due to the timing differences between replicate handling.

- Supplementary figures 3 through 6 show scatterplot matrices for each of the four clusters that resulted from hierarchical clustering analysis of the significant genes in the iron-metabolism soybean dataset (first shown in Figure 2 of the main paper). The data comes from the iron-metabolism soybean study, and have been filtered and standardized (Moran Lauter and Graham, 2016). Notice we see the same trends as we did in Figure 2 of the main paper.

- Supplementary figures 7 through 10 are the same scatterpot matrices as Figures 15-18 in the closing case study from the main paper, only now the data is *not* standardized. In general, we see that the genes that were called DEGs in both forms of normalization (purple and orange) have the expected differential expression profiles in the scatterplot matrices, deviating from the $x=y$ line in the treatment scatterplots in the anticipated direction (Supplementary figures 7 and 8). We also see that the genes that were removed with TMM normalization (red) do not show DEG patterns in the scatterplot matrices, as they barely deviate from the $x=y$ line in the treatment scatterplots (Supplementary figure 9. All three of these gene subsets appear as predicted. However, perhaps surprisingly, the genes that were added with TMM normalization (pink) appear similarly to the genes that were removed with TMM normalization (red) (Supplementary figure 10). We would expect the pink genes to deviate more from the $x=y$ line and demonstrate DEG patterns more than the red genes, but this was not observed. We solved this visualization problem in the main paper by using standardization techniques (Figures 15-18).

- Supplementary figures 11 through 14 are the same litre plots as Figures 19-22 in the closing case study from the main paper, only now the data is *not* standardized. Overall, we see that the example genes that were called DEGs in both forms of normalization (purple and orange) have the expected profiles in the litre plots, deviating as concentrated bundles away from the $x=y$ line (Supplementary figures 11 and 12). We also see that the example genes that were removed with TMM normalization (red) do not show DEG patterns in the litre plots, barely deviating from the $x=y$ and/or showing wide dispersion reflecting inconsistent replicates (Supplementary figure 13). All three of these gene subsets appear as predicted. However, perhaps surprisingly, the genes that were added with TMM normalization (pink) appear similarly to the genes that were removed with TMM normalization (red) (Supplementary figure 14). We would expect the pink genes to show DEG patterns (at least more so than the red genes), but this was not observed. We solved this visualization problem in the main paper by using standardization techniques (Figures 19-22).

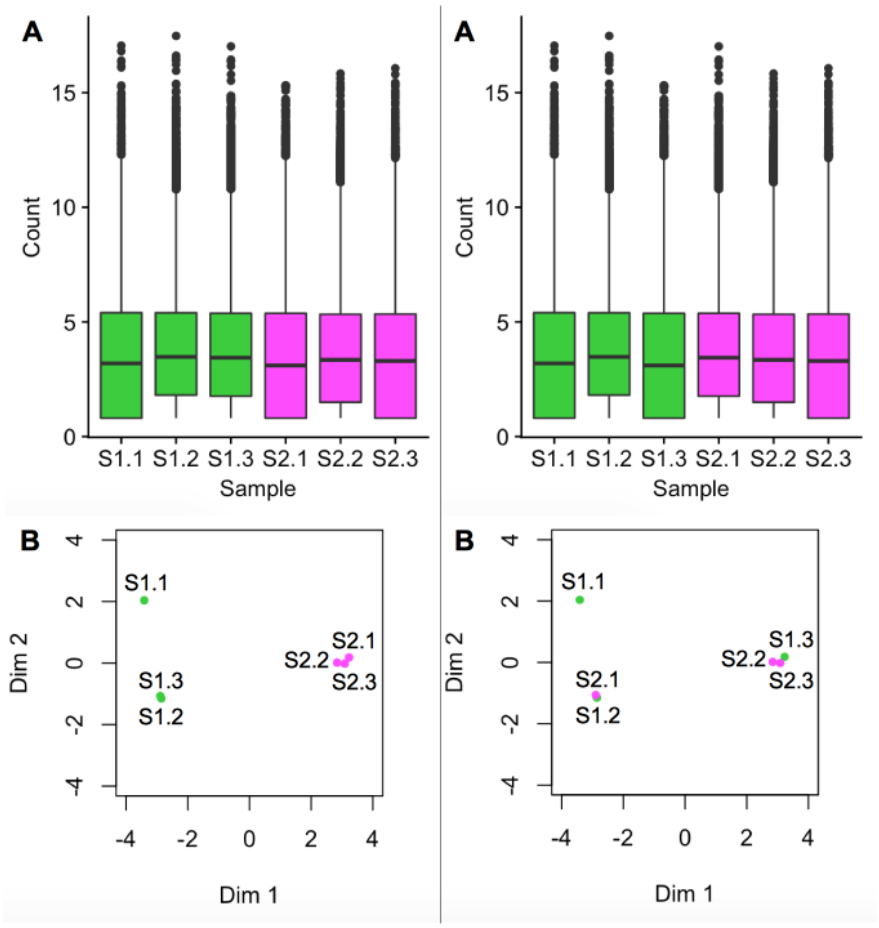

Figure 1: Boxplots and MDS plots are popular plotting tools for RNA-seq analysis. This figure shows these traditional visualization methods applied to the soybean cotyledon data before sample switching (left half) and after sample switching (right half) (Brown and Hudson, 2015). We cannot suspect from 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 less-popular plots that provide gene-level resolution like with the scatterplot matrix from Figure 5 in the main paper.

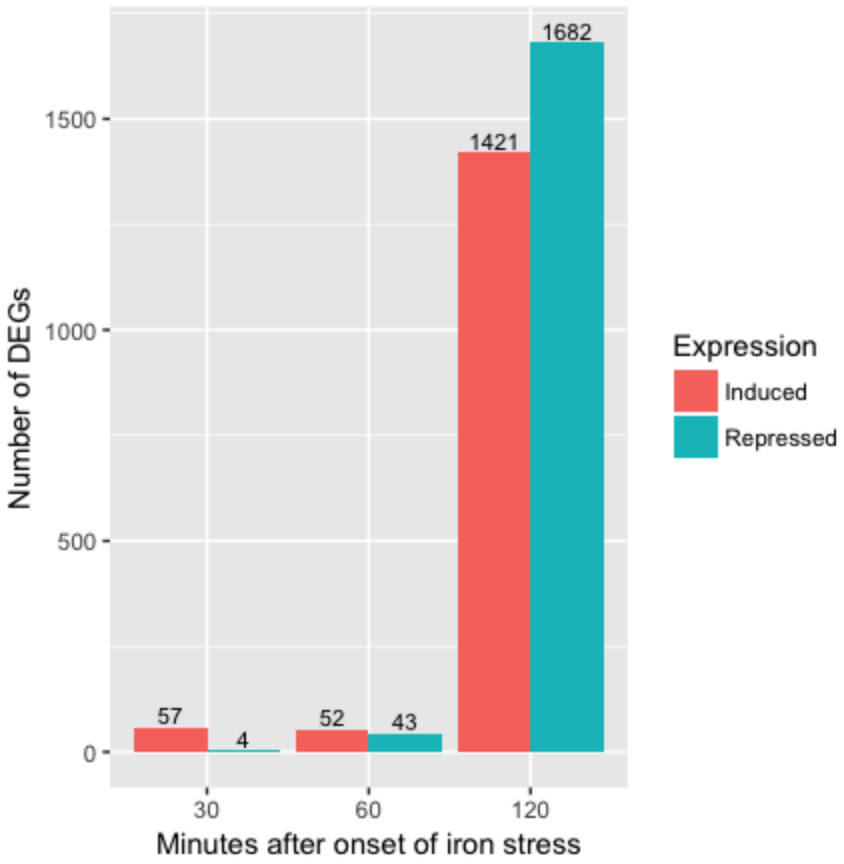

Figure 2: The authors of the soybean iron metabolism study (Moran Lauter and Graham, 2016) determined the DEGs across three times points (30 minutes, 60 minutes, and 120 minutes) in the leaves after onset of iron sufficient and deficient hydroponic conditions. They used the same researcher to collect the samples in succession. One major finding from their study was a vaste change in gene expression responses between these three time points. As a result, the streak observed in the scatterplot matrix containing the subset of data at the 120 minute time point (Figure 6 in the main paper) may be due to the timing differences between replicate handling.

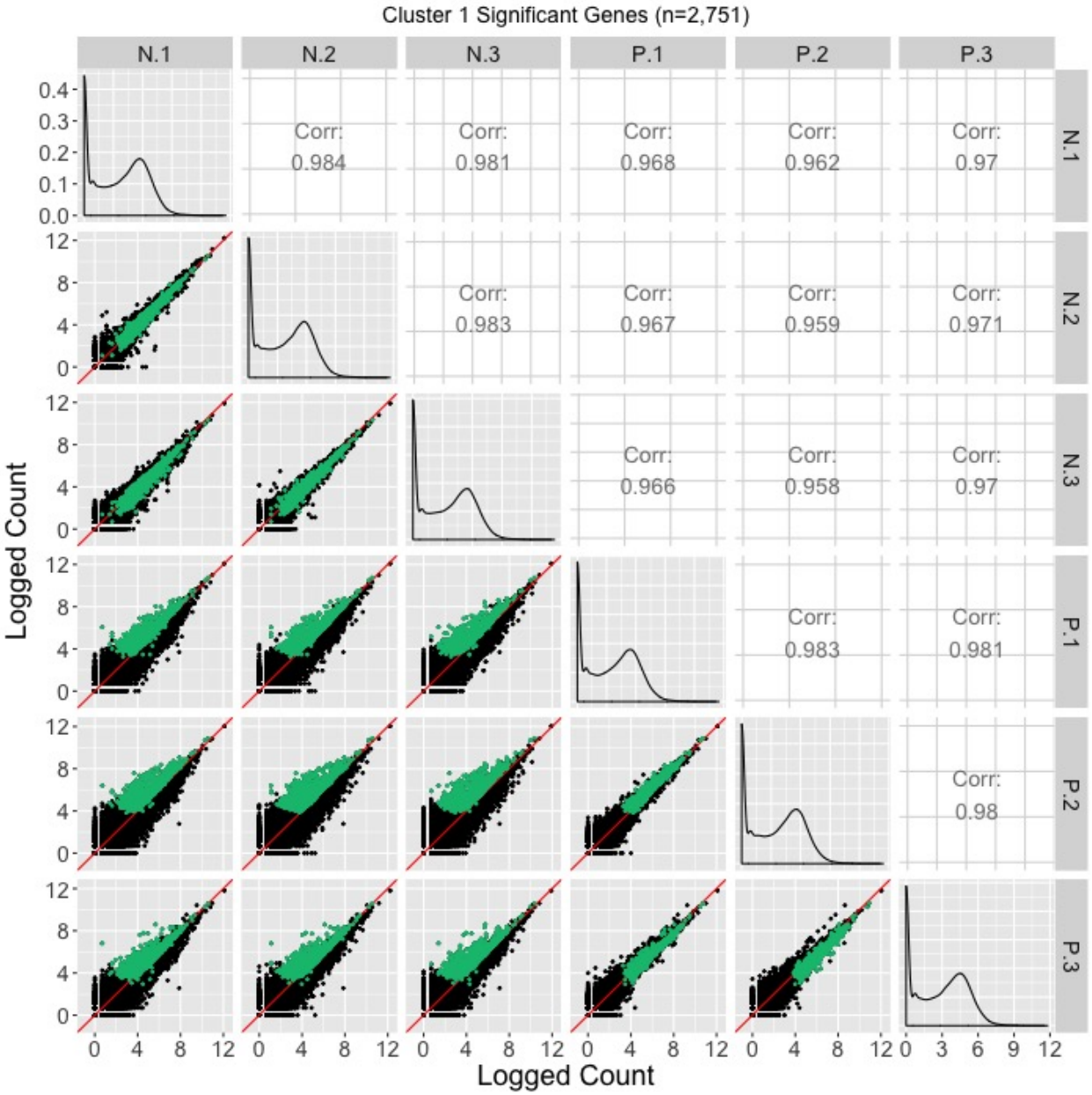

Figure 3: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2,751 significant genes in Cluster 1 after performing a hierarchical clustering analysis with a cluster size of four (Figure 2 in the main paper). These significant genes are overlaid in green on 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 DEGs 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 main 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. This is why we should also view these genes individually in litre plots (Figure 8 A and B of the main paper).

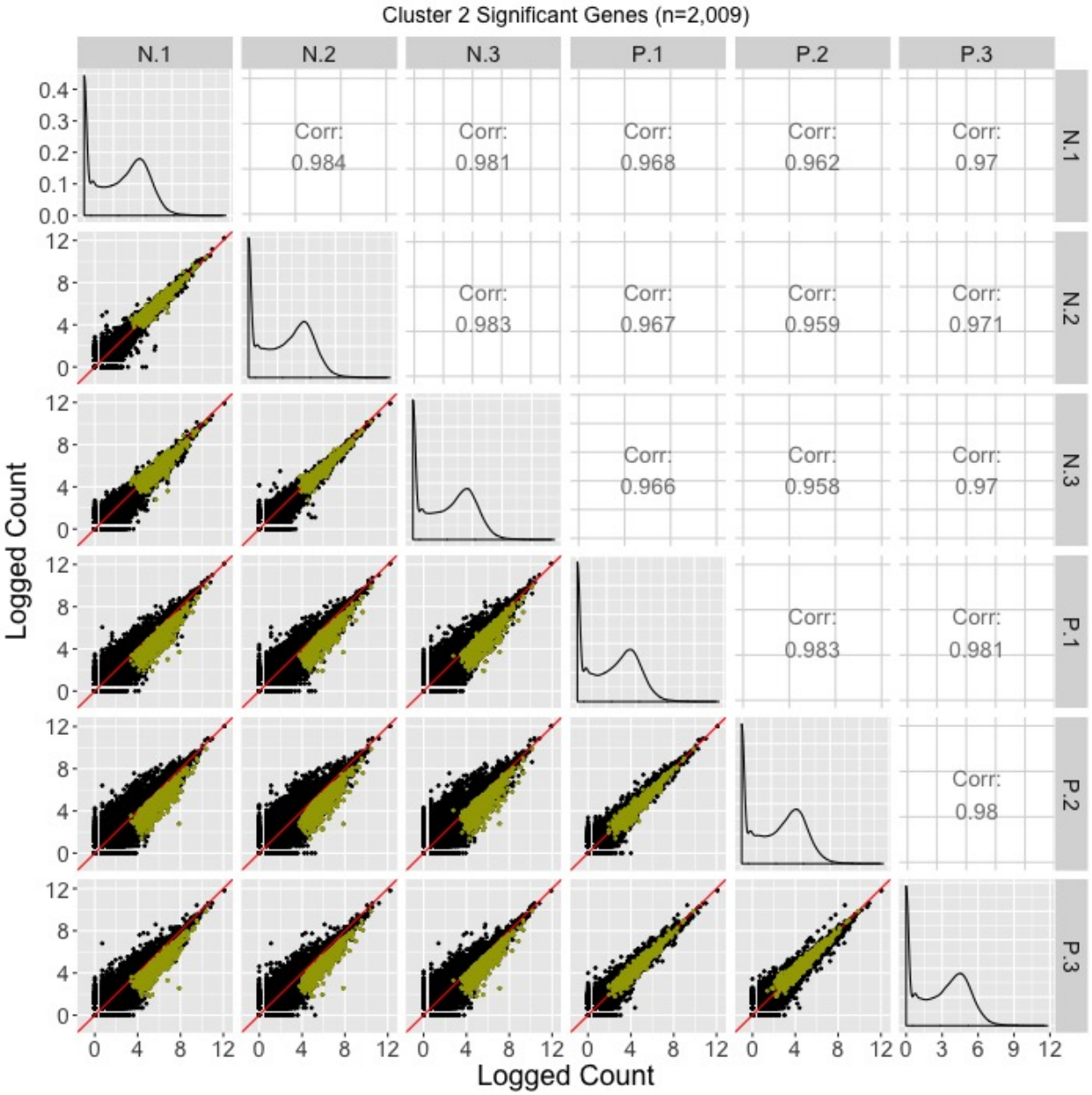

Figure 4: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 2,009 significant genes in Cluster 2 after performing a hierarchical clustering analysis with a cluster size of four (Figure 2 in the main paper). These significant genes are overlaid in dark yellow on the scatterplot matrix. They follow the expected patterns of differential expression with most dark yellow 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 dark yellow 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 main 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. This is why we might also view these genes individually in litre plots (Figure 8 C and D in the main paper).

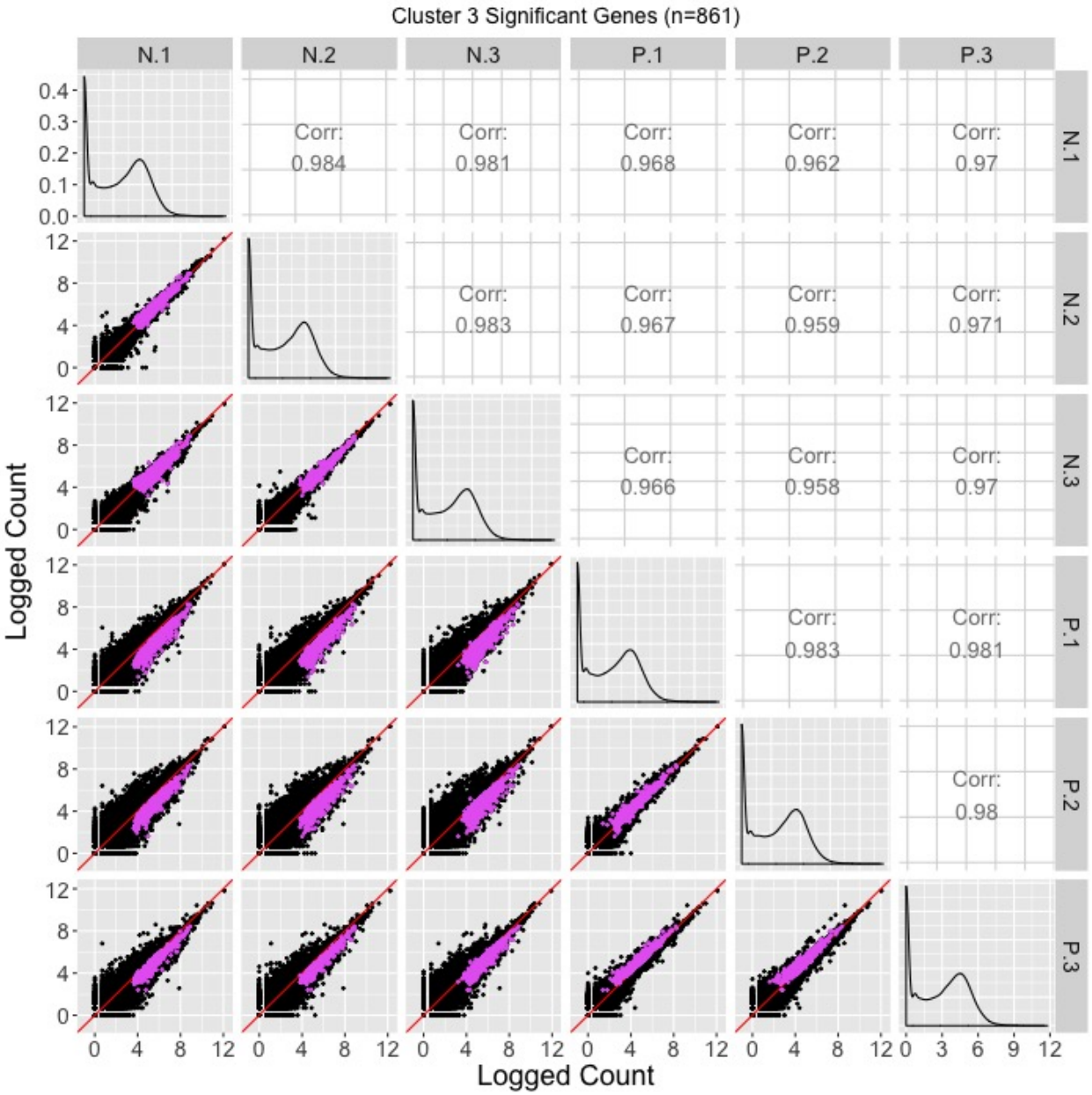

Figure 5: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 861 significant genes in Cluster 3 after performing a hierarchical clustering analysis with a cluster size of four (Figure 2 in the main paper). These significant genes are overlaid in pink on 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 shows slightly higher expression in P.3, and the scatterplot between replicates P.2 and P.3 also shows 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 in the main 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. This is why we might also view these genes individually in litre plots (Figure 8 E and F in the main paper).

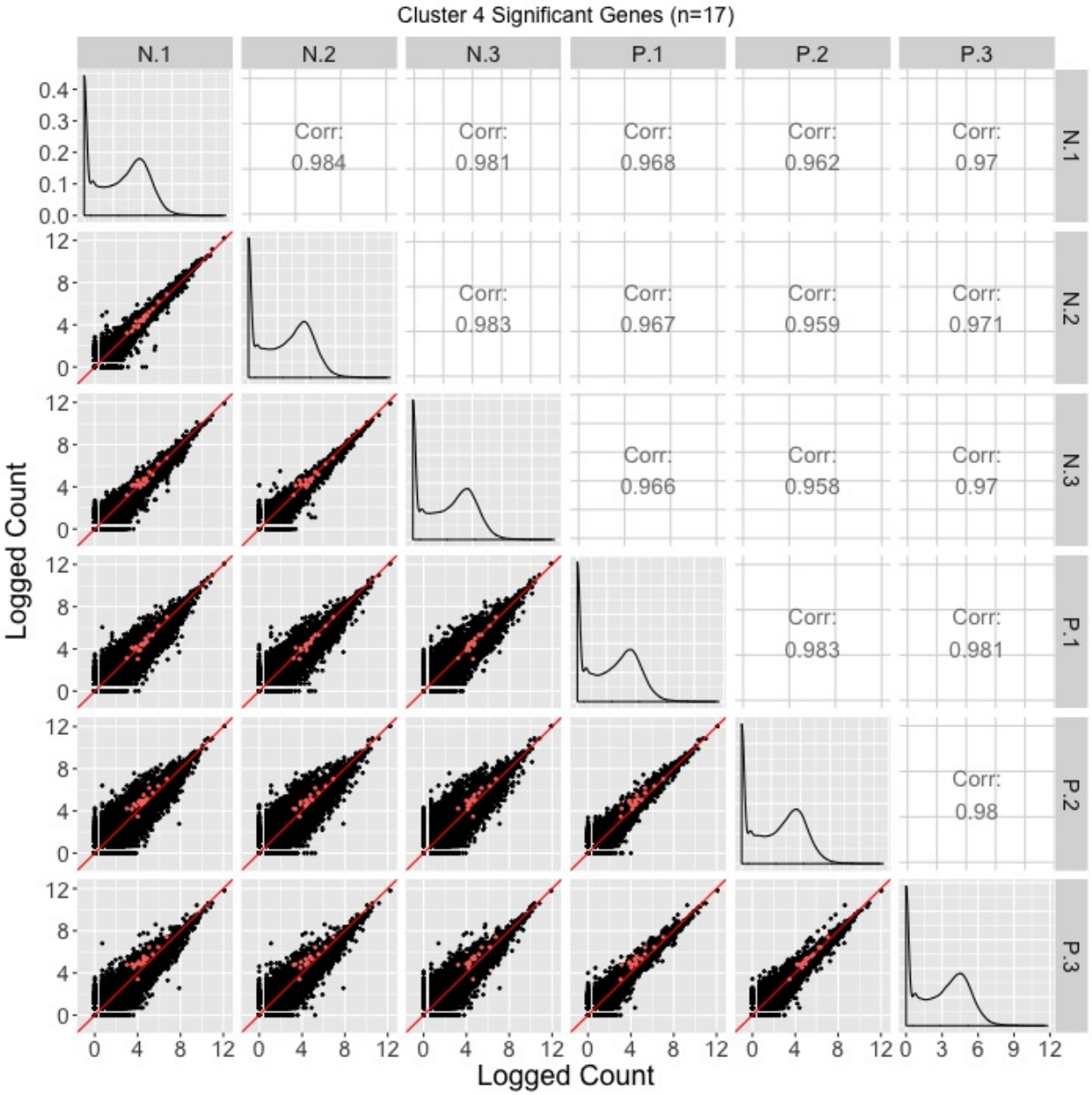

Figure 6: Example of using a scatterplot matrix to assess DEG calls from a model in the iron-metabolism soybean dataset. There were 17 significant genes in Cluster 4 after performing a hierarchical clustering analysis with a cluster size of four (Figure 2 in the main paper). These significant genes are overlaid in orange on 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 orange 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 orange points seems to indicate an underexpression of the P.1 replicate. We similarly saw somewhat messy looking DEG calls in Cluster 4 in the form of parallel coordinate plots (Figure 2 of the main paper) and litre plots (Figure 8 G and H of the main paper).

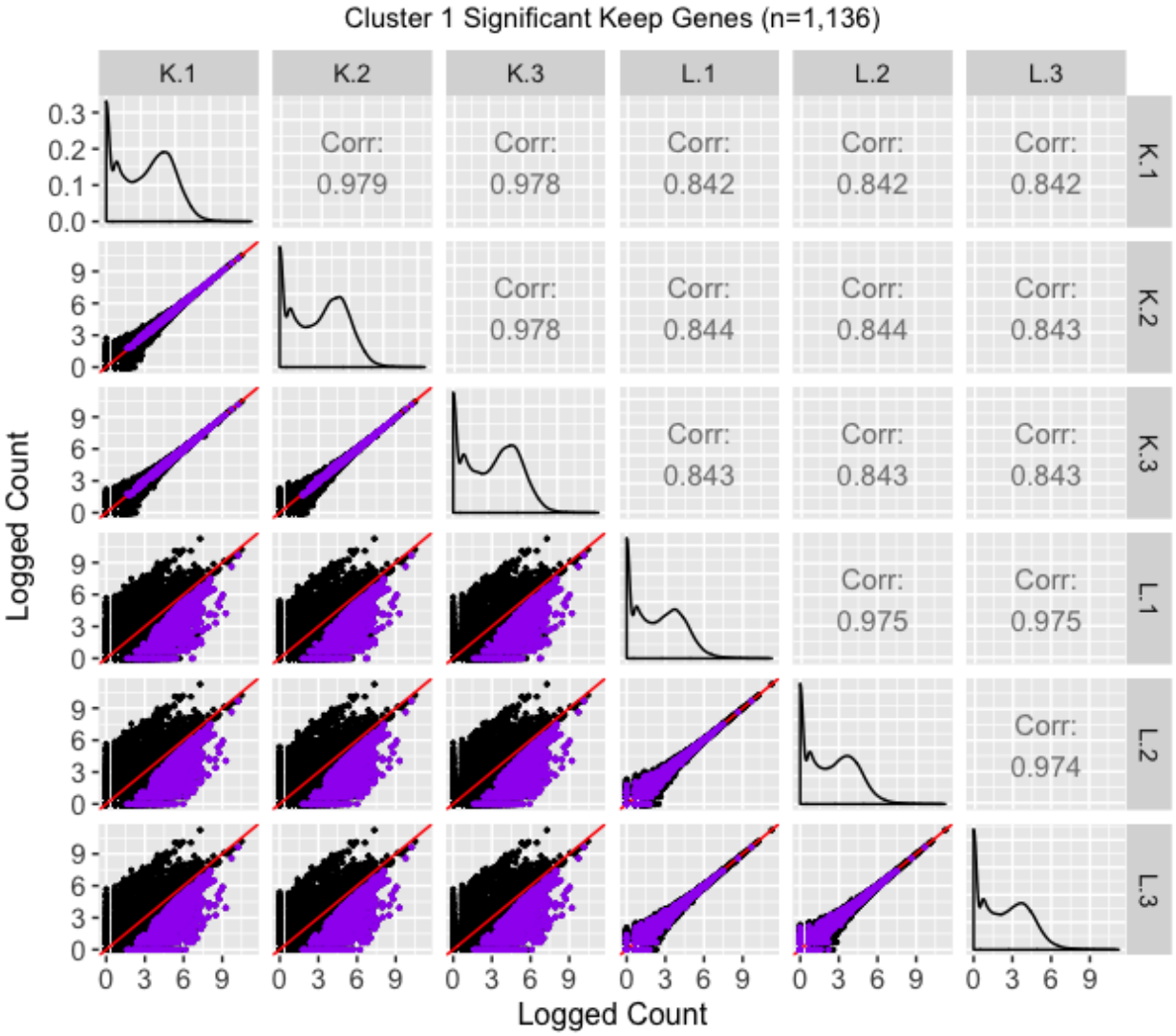

Figure 7: Scatterplot matrix of the 1,136 genes that were in the first cluster (Figure 11 of the main paper) from genes that remained as kidney-specific DEGs even after TMM normalization. With this scatterplot matrix, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

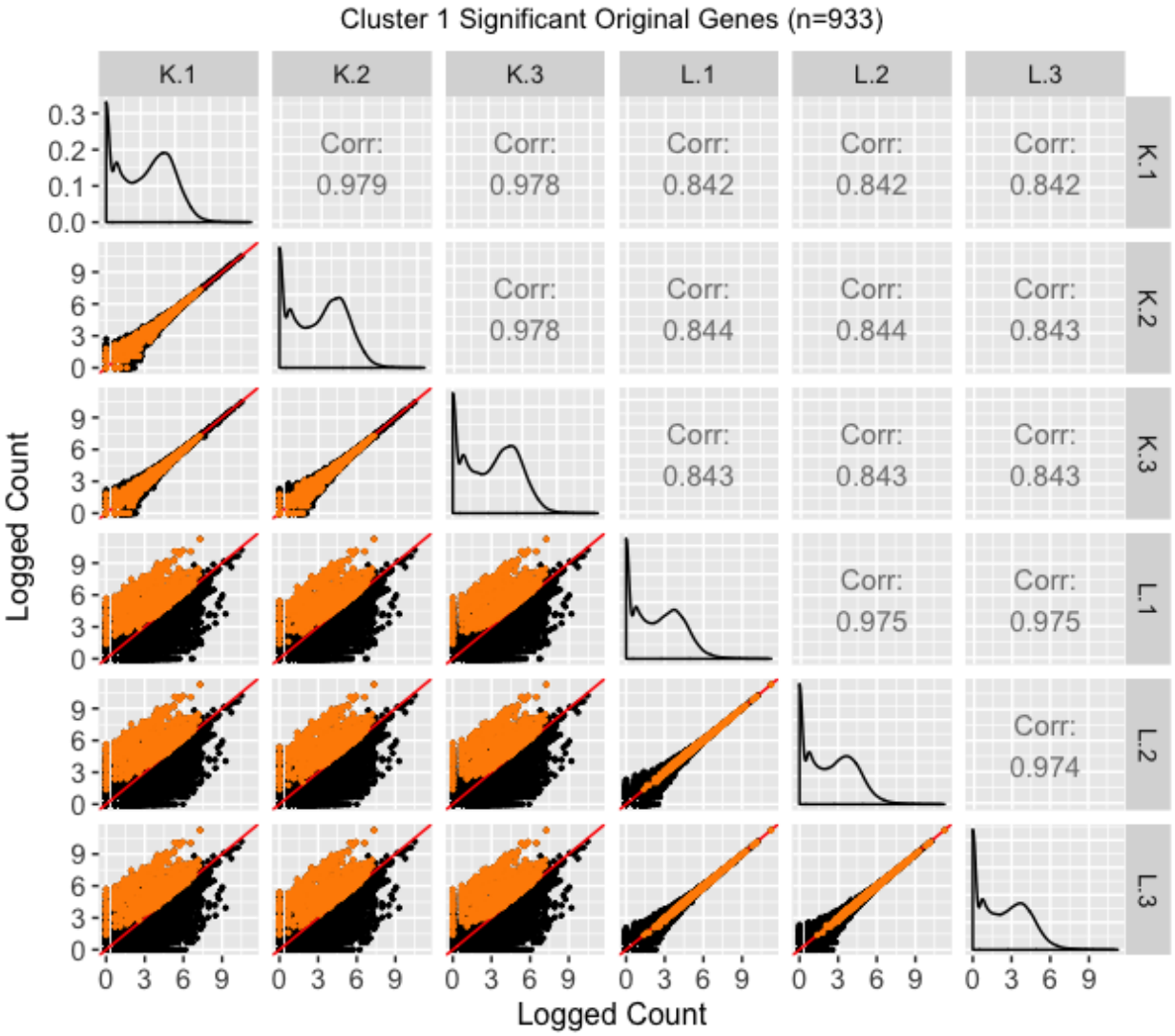

Figure 8: Scatterplot matrix of the 933 genes that were in the first cluster (Figure 12 of the main paper) from genes that remained as liver-specific DEGs even after TMM normalization. With this scatterplot matrix, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

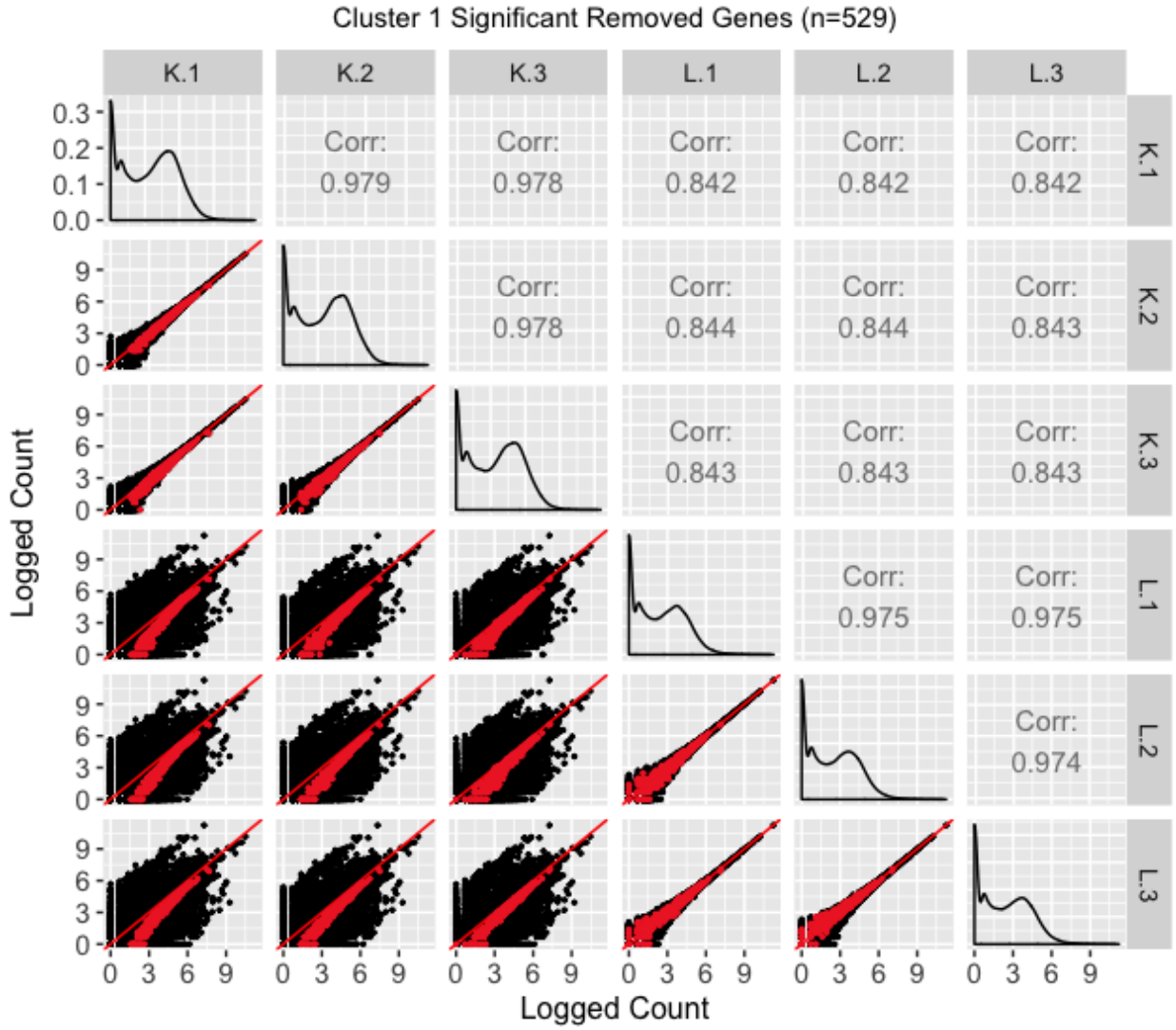

Figure 9: Scatterplot matrix of the 529 genes that were in the first cluster (Figure 13 of the main paper) from genes that no longer remained as kidney-specific DEGs after TMM normalization. With this scatterplot matrix, we verify from an additional perspective that these genes do *not* demonstrate the 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.

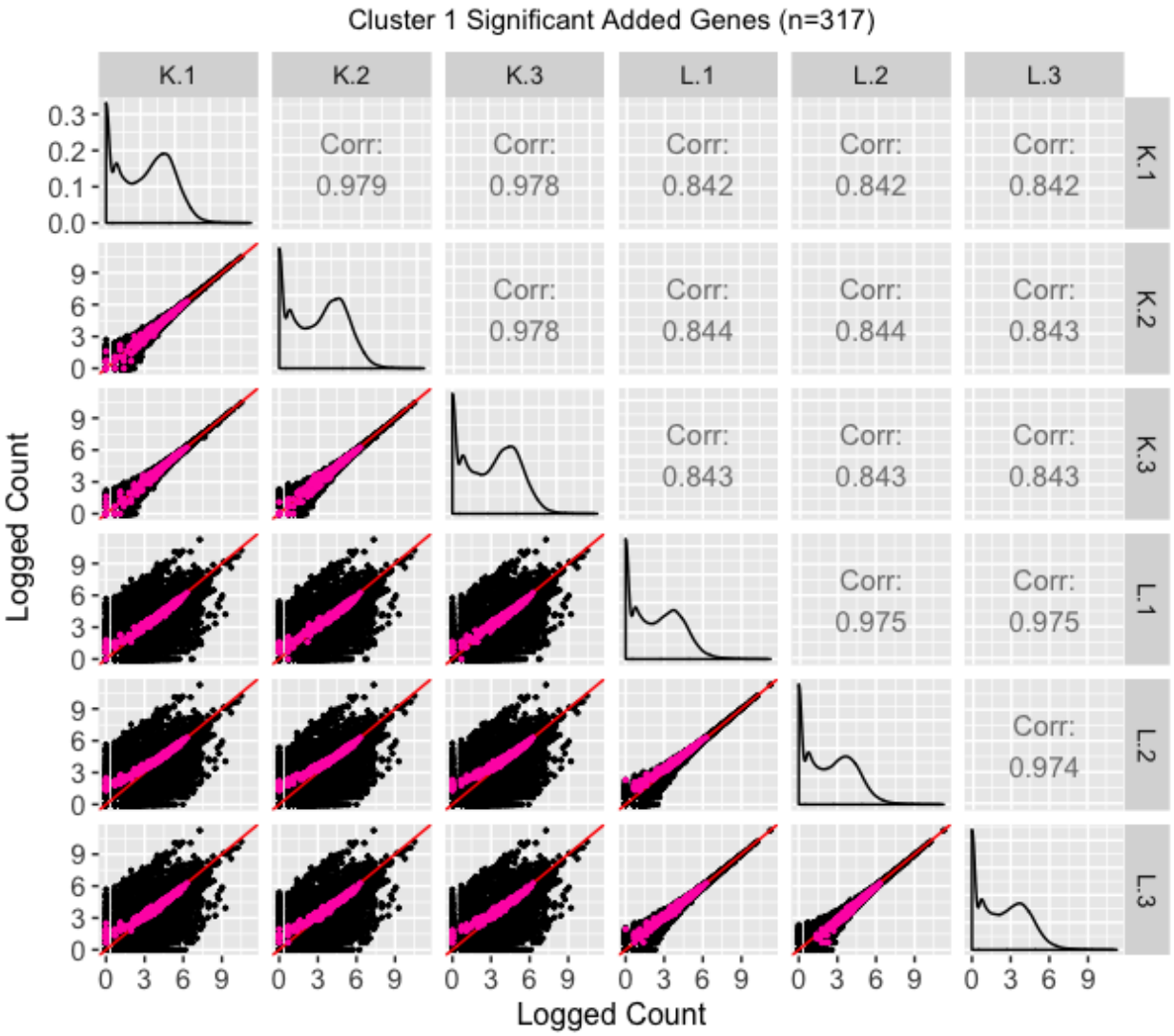

Figure 10: Scatterplot matrix of the 317 genes that were in the first cluster (Figure 14 of the main paper) from genes that were *added* as liver-specific DEGs after TMM normalization. With this scatterplot matrix, we see that the genes do *not* demonstrate the expected patterns of DEGs too strongly (they do not deviate much from the x=y line in the treatment scatterplots). In fact, these pink genes appear similarly to what we saw from the scatterplot matrix of the red genes (Supplementary figure 9). This is somewhat of a surprise, given that the pink genes were *added* by TMM normalization, while the red genes were *removed* by TMM normalization. Stated differently, we would expect the pink genes to appear more like differentially expressed genes if TMM normalization is appropriate, but we could not confirm this expectation. We solved this problem by standardizing the dataset as is shown in Figure 18 of the main paper.

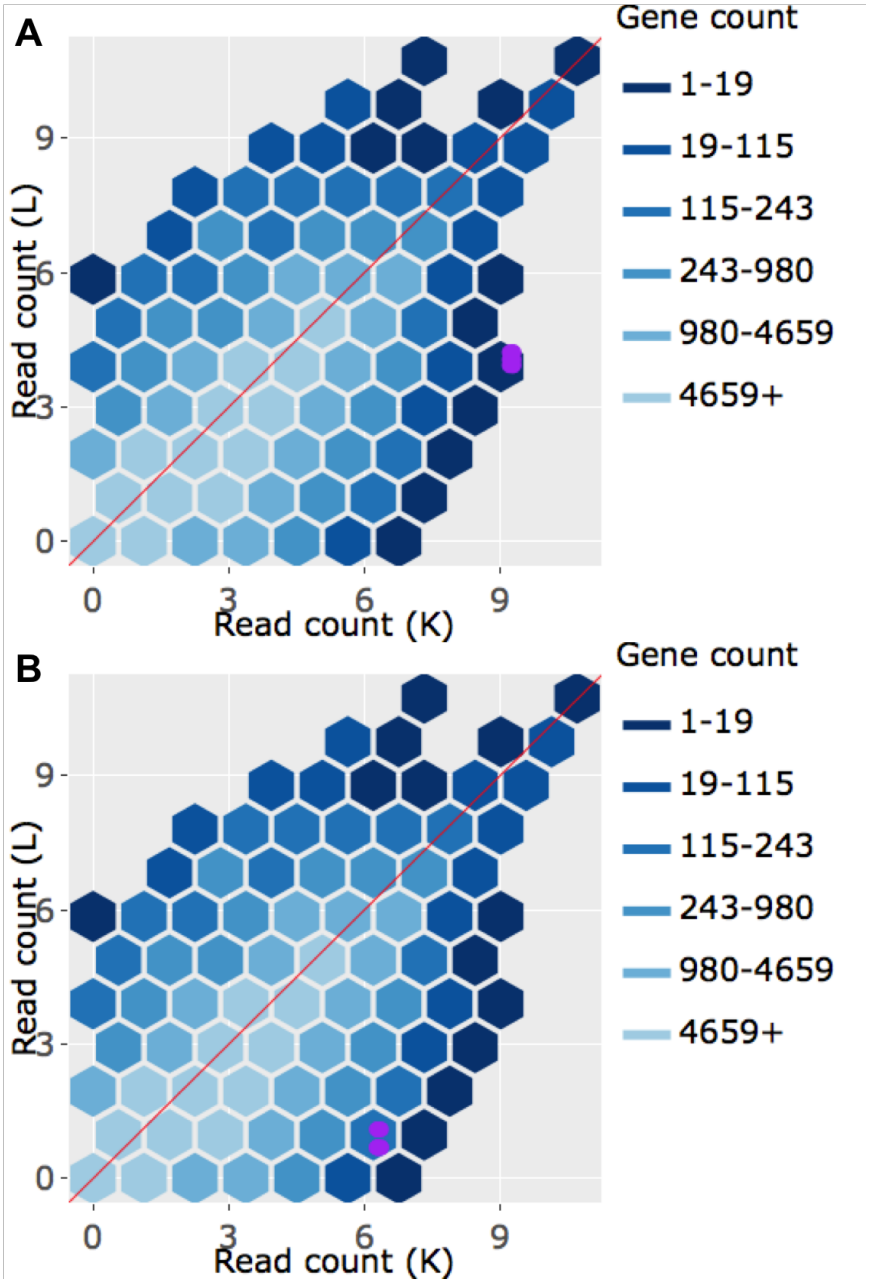

Figure 11: Example litre plots from the 1,136 genes that were in the first cluster (Figure 11 of the main paper) of genes that remained as kidney-specific DEGs even after TMM normalization. With these litre plots, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

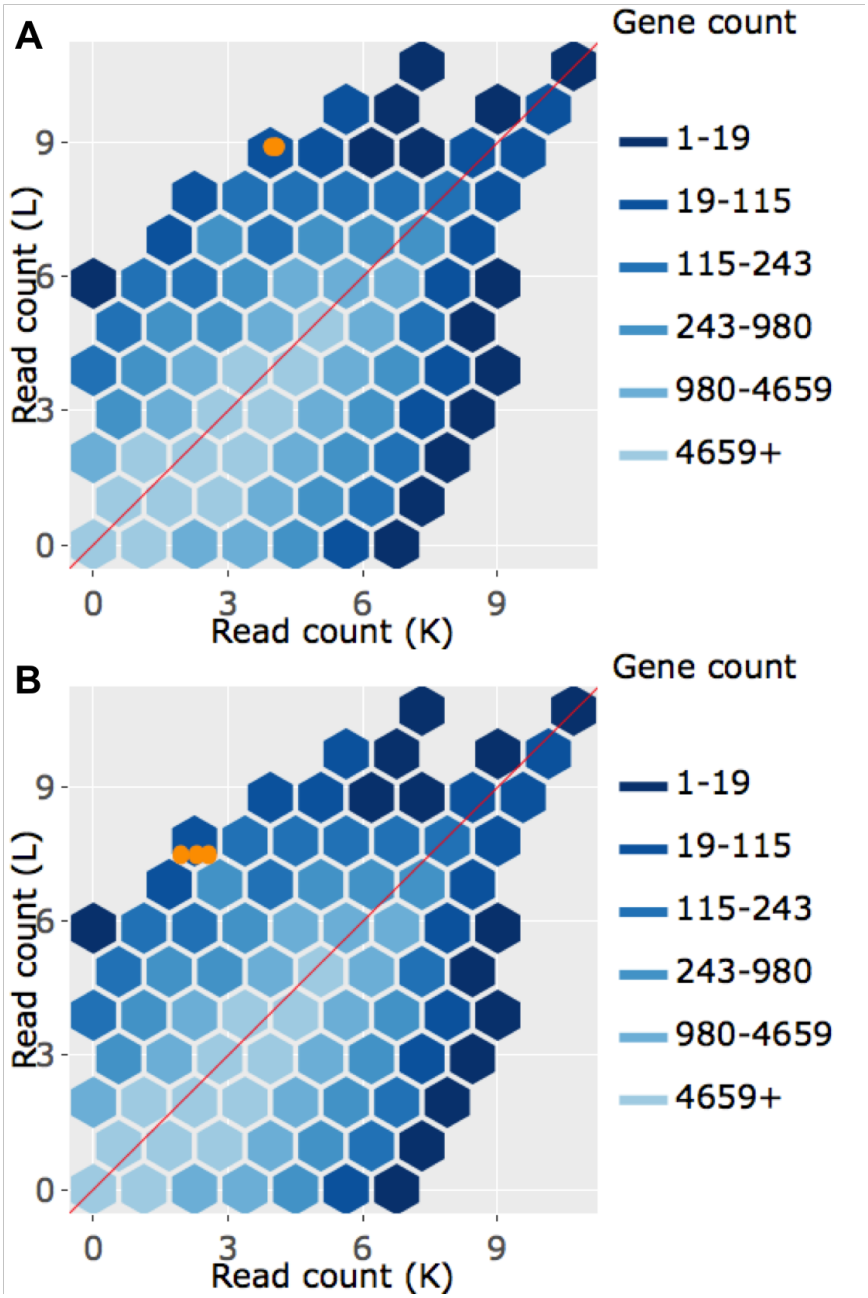

Figure 12: Example litre plots from the 933 genes that were in the first cluster (Figure 12 of the main paper) from genes that remained as liver-specific DEGs even after TMM normalization. With these litre plots, we verify from an additional perspective that these genes demonstrate the expected patterns of DEGs.

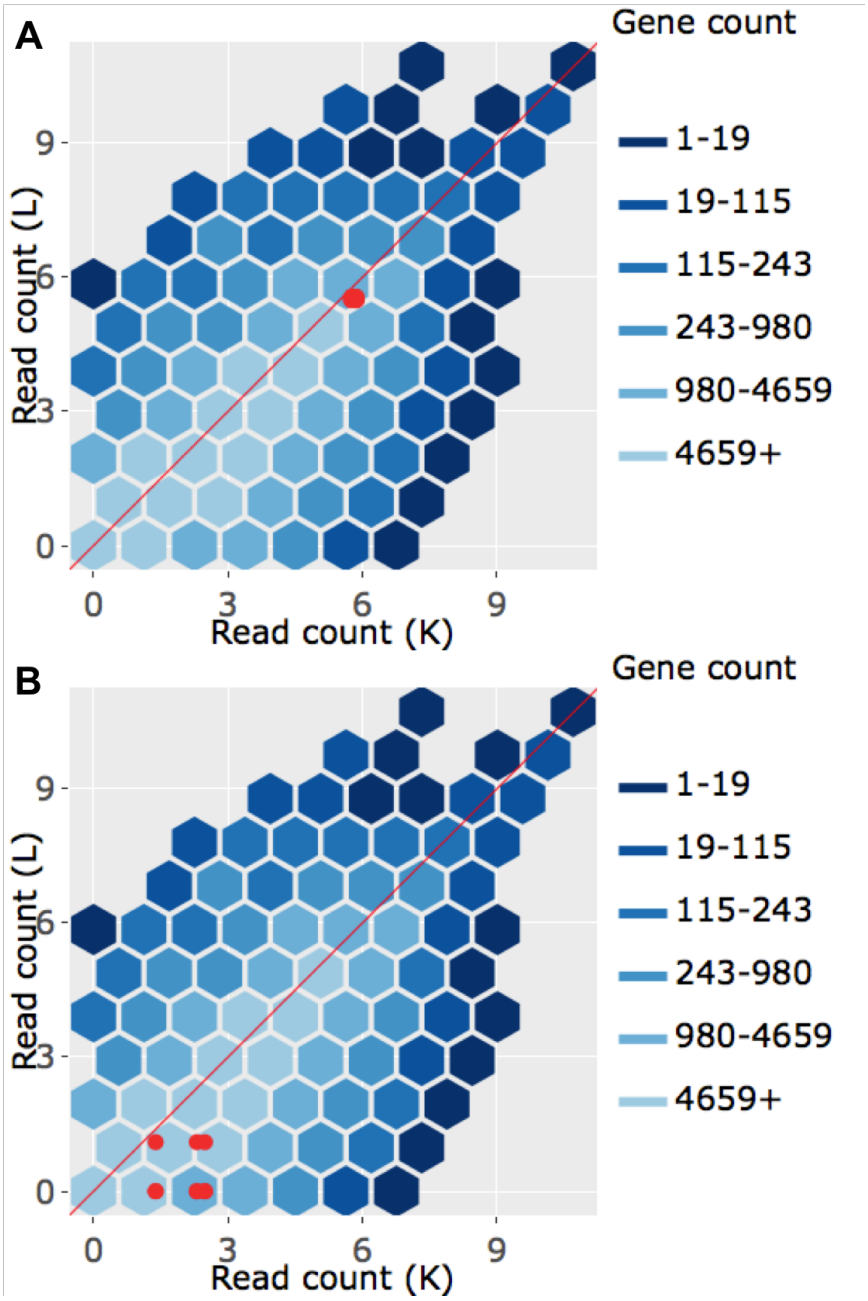

Figure 13: Example litre plots from the 529 genes that were in the first cluster (Figure 13 of the main paper) of genes that no longer remained as kidney-specific DEGs after TMM normalization. With these litre plots, we verify from an additional perspective that these genes do not demonstrate the expected patterns of DEGs. This provides additional evidence that TMM normalization removing these genes from DEG status may be valid.

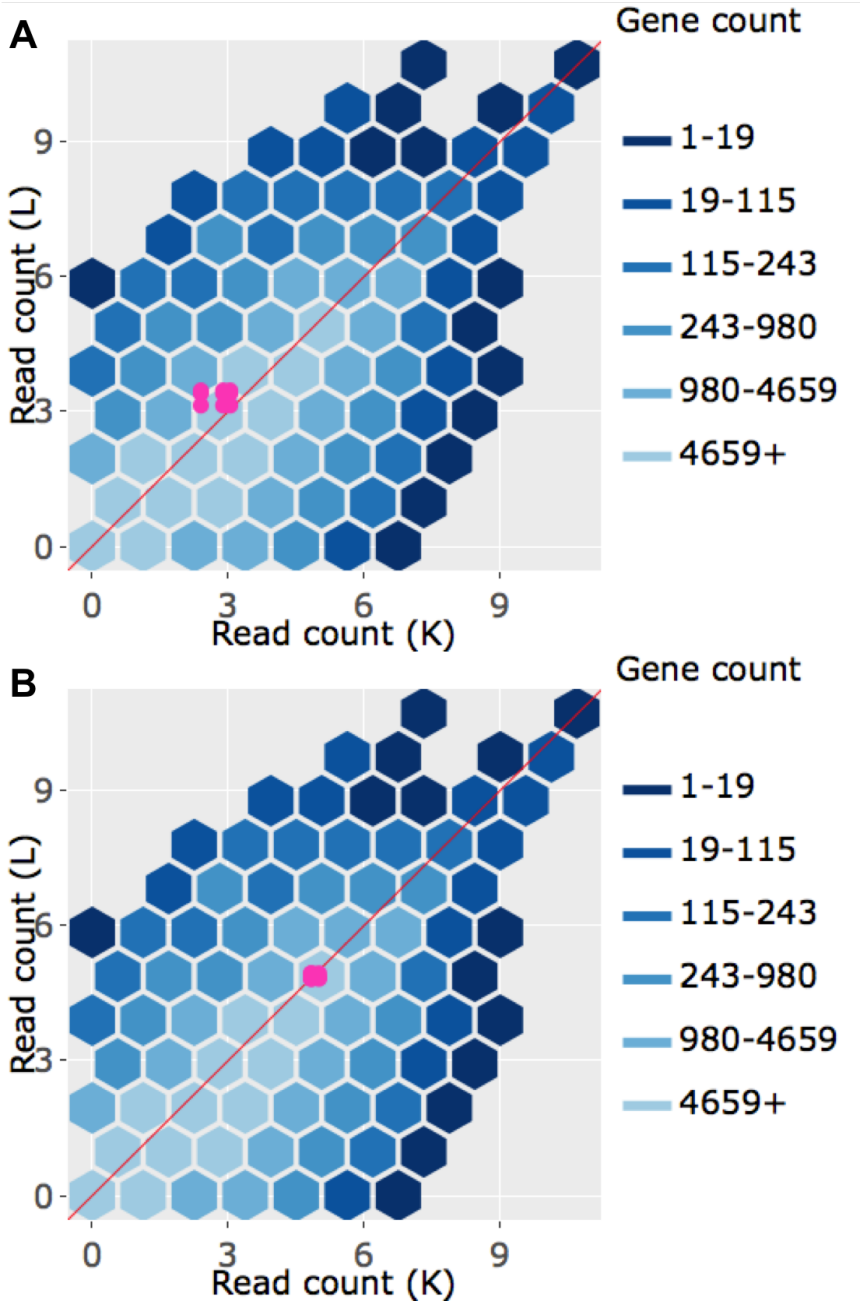

Figure 14: Example litre plots from the 317 genes that were in the first cluster (Figure 14 of the main paper) from genes that were *added* as liver-specific DEGs after TMM normalization. With these litre plots, we see that the genes do *not* demonstrate the expected patterns of DEGs in a trustworthy manner. In fact, these pink genes appear similarly to what we saw from the example litre plots of the red genes (Supplementary figure 13). This is somewhat of a surprise, given that the pink genes were *added* by TMM normalization, while the red genes were *removed* by TMM normalization. Stated differently, we would expect the pink genes to appear more like differentially expressed genes if TMM normalization is appropriate, but we could not confirm this expectation. We resolved this problem by standardizing the dataset as is shown in Figure 22 of the main paper.