This is the supporting material to our paper on “Depth Normalization of Small RNA Sequencing: Using Data and Biology to Select a Suitable Method”. We show additional results of the DANA approach for assessing the performance of normalization methods in microRNA sequencing data using (1) the paired MSK data sets, (2) the partially-paired TCGA-UCEC data sets and (3) the combined TCGA-BRCA and TCGA-UCS data set. All results were generated in R version 4.0.2 and the code, data, and results are available on GitHub at https://github.com/LXQin/DANA-paper-supplementary-materials.

1 MSK Data

We further study the paired MSK data sets for the same set of tumor samples for which a detailed description of the data was previously reported in Qin et al. (2020). Recall that one data set (that is, the benchmark data set) was collected using (1) uniform handling to minimize data artifacts and (2) balanced sample-to-library-assignment (via the use of blocking and randomization) to balance any residual artifacts with the tumor groups under comparison. For the same set of samples, a second data set (that is, the test data set) was collected without making use of such a careful study design, resulting in unwanted depth variations. Using this pair of data sets, we empirically examine the DANA approach by evaluating its sensitivity for different control marker choices, assessing different method choices for partial correlation estimation, and comparing control-marker correlations in negative and positive controls between the two data sets.

1.1 Cutoff choices for defining control makers

We empirically study the sensitivity of the DANA method for three different cutoff choices. The three chosen configurations are (a) $[\ell^-, u^-] = [1, 4]$ and $[\ell^+, \infty) = [64, \infty)$; (b) $[\ell^-, u^-] = [2, 10]$ and $[\ell^+, \infty) = [128, \infty)$, which corresponds to the control definition in the main paper; and (c) $[\ell^-, u^-] = [3, 16]$ and $[\ell^+, \infty) = [256, \infty)$. The number of negative controls are (a): 92, (b): 102, and (c): 103, and the number of positive controls are (a): 150, (b): 115, and (c): 77. For each configuration, we followed the guidelines presented in the main paper and first selected $\ell^-$ such that selected poorly-expressed miRNAs show at least mild expression and then selected $u^-$ and $\ell^+$ such that each control group consists of a sufficient number of markers (Fig. 1, top and middle panel).
We observe that the cutoff choices only mildly affect the DANA results (Fig. 1, bottom). For all configurations under study, we observe very similar relative DANA metrics for all considered normalization methods; where, generally, DESeq and TMM offer the best trade-off between high reduction of handling effects and high preservation of biological signals. Hence, we conclude that DANA is not sensitive to the cutoff choices for defining the control markers.

For all further results, we use the same definition of negative controls ([\ell^-, u^-] = [2, 10]) and positive controls ([\ell^+, \infty] = [128, \infty]) as in the main paper.

### 1.2 Method consideration and tuning parameter calibration for the estimation of partial correlations in positive controls

For positive controls, we examine different approaches for the estimation of partial correlations. We compare the estimated partial correlation matrices and the DANA results for the test data us-
ing (1) the neighborhood selection by Meinshausen and Bühlmann (2006) (MB), (2) the glasso method (Friedman et al., 2008), and (3) the FastGGM method (Ren et al., 2015). The MB method requires calibration of a tuning parameter, for which we consider Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), 10-fold Cross-Validation (CV), and Adaptive Validation (AV) (Chichignoud et al., 2016). We, further, consider tuning parameter calibration using the Stability Approach for Regularization Selection (StARS) and the Rotation Information Criterion (RIC) for MB and glasso, as implemented in the R huge package (Zhao et al., 2012), and also the default tuning parameter calibration for the FastGGM method (Wang et al., 2016).

We observe relatively similar correlation patterns for the un-normalized MSK test data set using MB or glasso for all tuning parameter calibration methods (Fig. 2). The correlations estimated using MB, CV and AV show the most conservative tuning, while RIC shows the most liberal one with only few non-zero partial correlations. Correlations estimated using glasso show similar patterns to MB; however, overall correlation strengths are all low in comparison to MB with only a few high partial correlations. This indicates that glasso may be more suitable for graphical model estimation than partial correlation estimation and, hence, we do not use it as the default method for DANA. FastGGM estimated a dense partial correlation matrix, and, hence, may not be suitable for our application; this may be an artifact to its tuning parameter calibration and will be subject to future study.

Finally, we compare estimated partial correlations with Pearson and Spearman correlation. In contrast to the partial correlation matrices estimated using MB or glasso, Pearson and Spearman correlation matrices estimate fully populated correlation matrices. However, in positive controls, we intend to capture only the direct co-expression relation for each pair of markers, removing any spurious correlation due to co-expression with other positive controls. Hence, partial correlations are more suitable than marginal correlations.

We apply our DANA method to the MSK test data using all of the aforementioned precision estimation methods (Fig. 3). We observe that using the MB method the relative performance of each normalization method stays the same regardless of the tuning parameter calibration used for MB. Similarly, comparing the relative $cc^+$ metrics of normalization methods, glasso yields similar results to MB for both RIC and StARS. Indeed, we observe very high correlation of the DANA $cc^+$ metrics across all MB methods, high correlation of MB with glasso+RIC, but only moderate correlation between MB and glasso+StARS (Fig. 4). Even though FastGGM estimated a dense correlation matrix, we observe high to very high correlation with MB depending on the tuning parameter calibration used for MB.

We reason that it is essential for DANA to accurately estimate the relative strengths of within-cluster partial correlations, which are used for computing the $cc^+$ metric, rather than correctly estimating the sparsity of partial correlations; hence, MB and FastGGM are more suitable for the DANA method than glasso. We decide to use the MB method as a default for the DANA method since it correctly estimates a sparse partial correlation matrix. We use BIC tuning as its computation is fast and simple.

1.3 Partial correlations of positive controls in the test data

For positive controls, all within-cluster partial correlations in the test data are strictly positive as expected regardless of normalization (Fig. 5). Furthermore we observe a high abundance of strong positive partial correlations for clustered markers as we expected based on biological evidence reported in the literature on polycistronic clusters. Compared to the benchmark data (see the main paper), the test data has fewer positive within-cluster correlations and more positive, off-cluster correlations, both of which likely resulted from excessive handling effects. Normalization, regardless of the method, alleviated the off-cluster correlations in terms of both the number and strength; depending on the method, normalization can either retain or reduce within-cluster correlations, which is signified by their before-versus-after-normalization concordance. This concordance strongly varies with the used normalization method (Fig. 6). TMM ($cc^+ = 0.96$), DESeq ($0.95$), PoissonSeq ($0.96$), TC ($0.96$), RUVg ($0.96$), and RUVs ($0.92$) all show very high concordance as measured by the metric $cc^+$. RUVR ($0.81$), UQ ($0.70$), and Med ($0.68$) fail to achieve such high $cc^+$ with a much bigger spread of corresponding correlations in the partial correlation scatter plots compared to the aforementioned methods.
Figure 2: Estimated correlations in positive controls for the un-normalized MSK test data. The correlation matrix is displayed using heatmaps for each correlation estimation method under study. The three top rows show estimated partial correlations and the bottom row shows estimated marginal Pearson (bottom left) or Spearman (bottom center) correlations, respectively. In each heatmap, the upper triangular matrix shows the correlation strength and the lower triangular matrix indicates miRNA polycistronic clustering.
Figure 3: Scatter plots of the two summary metrics for the MSK test data using varying partial correlation estimation methods for positive controls. Each point stands for a normalization method under study. Note that varying the partial correlation estimation method, only the $cc^+$ metric changes and the mscr− metric, which is computed using negative controls only, stays constant across methods.
1.4 Marginal correlations of negative controls in the benchmark and test data

For negative controls, we assess the level of inter-marker correlations in the test data before and after normalization, and compare these with that in the benchmark data (Fig. 7). The frequency of strong correlations in the test data is much higher compared to the benchmark data while, generally, most correlations in the benchmark data are weak (centered around 0.21). The variation of correlation strengths in the test data is much higher, and, in particular, stronger positive correlations are much more abundant compared to the benchmark data. This finding accords with our assumption that handling effects manifest themselves as excessively strong positive correlations. The level of correlation strengths and correlation variability after normalization vary among the different normalization methods with some methods leading to a stronger correlation reduction compared to others; here, RUVr stands out from the other methods and achieved the highest correlation reduction. Furthermore, note that the correlation histogram curves for TC, UQ, and Med are very similar due to the similarity of these methods.

For normalized benchmark data (Fig. 8), we observe that most normalization methods show very similar correlation variance. However, some methods (RUVg, TMM, DESeq, and QN) show better centering than the others. RUVs stands out, since it introduces some strong positive correlations and
Figure 5: Heatmaps of the estimated correlations among positive controls in the normalized MSK test data for each normalization method and for un-normalized MSK test data. In each heatmap, the upper triangular matrix shows the correlation strength and the lower triangular matrix indicates miRNA polycistronic clustering.
Figure 6: Scatter plots of partial correlations in positive controls before versus after normalization for each normalization method in the MSK test data. The corresponding concordance correlation coefficient between the shown partial correlations, which is measured by $cc^+$, is indicated in the lower right corner of each plot.
Figure 7: Histogram frequency curve for marginal correlations in negative controls for the un-normalized benchmark data, un-normalized test data, and test data normalized using each normalization method under study.

shows an overall unbalanced correlation pattern with a spike for small negative correlations. Hence, RUVs is not suitable for normalization of the benchmark data.
Figure 8: Histogram frequency curve for marginal correlations in negative controls for the un-normalized benchmark data, un-normalized test data, and benchmark data normalized using each normalization method under study.
2 TCGA-UCEC Data

2.1 Data Description

We constructed the two data sets so that they each contain 48 samples, 22 of which were of endometri-oid subtype (END) and 26 of serous subtype (SER), and so that these two data sets have 24 samples in common. The 48 samples of the first data set were processed in a single batch (batch number 228.63.0), which we refer to as the “single-batch” data. The second data set (that is the “mixed-batch”) is composed of 24 samples (11 END and 13 SER) from the first data set (batch 228.63.0) and 24 samples (11 END and 13 SER) from other batches; END samples are from the batches 104.85.0 (1 sample), 121.83.0 (1), 156.80.0 (1), 186.74.0 (2), 59.83.0 (2), 73.85.0 (2), 81.83.0 (1), and 92.87.0 (1); and SER samples are from the batches 110.75.0 (1), 137.81.0 (1), 143.82.0 (1), 156.80.0 (1), 168.78.0 (2), 178.78.0 (1), 186.74.0 (1), 201.70.0 (1), 324.54.0 (1), 381.43.0 (1), 49.86.0 (1), and 73.85.0 (1). We refer to the latter data set as the “mixed-batch” data. Sequencing counts are available for 1848 miRNAs in each sample.

2.2 Correlation in positive and negative controls

For negative controls, we assess the level of their inter-marker Pearson correlation in the mixed-batch data before and after normalization, and compare these with that in the single-batch data (Fig. 9). As for the MSK data sets, the frequency of strong correlations in the mixed-batch data is higher compared to the single-batch data while, generally, most correlations in the single-batch data are weak. This finding, again, accords with our assumption that handling effects manifest themselves as excessively strong positive correlations. Normalization of the mixed-batch data not only drastically reduces strong positive correlations but also shifts positive correlations towards zero for most normalization methods; except for TC, UQ, Med, and RUVs, positive correlations in negative controls are alleviated below the level of correlations in the single-batch data. Hence, these methods provide effective removal of handling effects in the handling-prone mixed-batch data.

For positive controls, we observe a high abundance of within-cluster correlations in the single-batch all of which are strictly positive as expected (Fig. 10). As we have reported in the comparison of the MSK benchmark and test data, the mixed-batch data has fewer positive within-cluster correlations and more off-cluster positive correlations compared to the single-batch data. Both likely, again, resulted from the added handling effects in the mixed-batch data. Normalization, regardless of the method, alleviated the off-cluster correlations in terms of both the number and strength. Finally, depending on the method, normalization can either retain or eliminate within-cluster correlations, which is signified by their before-versus-after-normalization concordance, that is, the $cc^+$ metric (Fig. 11).
Figure 9: Histogram frequency curve for marginal correlations in negative controls for the un-normalized single-batch data, un-normalized mixed-batch data, and normalized mixed-batch data using each normalization method under study.
Figure 10: Heatmaps of the estimated correlations among positive controls in the normalized TCGA-UCEC mixed-batch data for each normalization method as well as for un-normalized TCGA-UCEC single-batch (top left) and un-normalized mixed-batch (top center) data. In each heatmap, the upper triangular matrix shows the correlation strength and the lower triangular matrix indicates miRNA polycistronic clustering.
Figure 11: Scatter plots of partial correlations in positive controls before versus after normalization for each normalization method for the TCGA-UCEC mixed-batch data. The corresponding concordance correlation coefficient between the shown partial correlations, which is measured by $cc^+$, is indicated in the lower right corner of each plot.
Figure 12: Mean count histogram (left) and mean-standard deviation plot (right) for the TCGA-BRCA/UCS data. The ranges $[\ell^-, u^-] = [2, 5]$ for negative controls and $[\ell^+, \infty) = [100, \infty)$ for positive controls are indicated by blue and red vertical lines, respectively.

3 TCGA-BRCA and TCGA-UCS Data

3.1 Control marker definition

Recall that the TCGA-BRCA/TCGA-UCS data set combines 166 stage III and IV samples from the TCGA-BRCA project with all 57 samples from the TCGA-UCS project. This data set does not have an available benchmark. We selected the ranges $[\ell^-, u^-] = [2, 5]$ and $[\ell^+, \infty) = [100, \infty)$ based on our recommended data-driven graphical criteria (Fig. 12). The number of negative controls is $p^- = 91$ and the number of positive controls is $p^+ = 116$.

3.2 Correlation in positive and negative controls

For negative controls, we assess the level of their inter-marker correlation in the combined TCGA-BRCA and TCGA-UCS data set (Fig. 13). As for the MSK and TCGA-UCEC data sets, normalization drastically reduces strong positive correlations in negative controls and different normalization leads to different reduction of correlations and, hence, different reduction of handling effects, which is quantified by the $\text{mscr}^-$ metric. Note further that similar normalization methods lead to similar correlation distributions; for example, Med and UQ normalization show a near-identical distribution.

For positive controls, we again observe a high abundance of within-cluster correlations as expected for clustered miRNAs (Fig. 14). Normalization can alleviate off-cluster correlations in terms of both the number and strength. Finally, depending on the method, normalization can either retain or reduce within-cluster correlations, which is signified by their before-versus-after-normalization concordance (Fig. 15).
Figure 13: Histogram frequency curve for marginal correlations in negative controls for the un-normalized TCGA-BRCA/UCS data and normalized TCGA-BRCA/UCS data for each normalization method under study.
Figure 14: Heatmaps of the estimated correlations among positive controls of the TCGA-BRCA/UCS data set before and after normalization. In each heatmap, the upper triangular matrix shows the correlation strength and the lower triangular matrix indicates miRNA polycistronic clustering.
Figure 15: Scatter plots of partial correlations in positive controls before versus after normalization for each normalization method for the TCGA-BRCA/UCS data. The corresponding concordance correlation coefficient between the shown partial correlations, which is measured by \( cc^+ \), is indicated in the lower right corner of each plot.
References

Chichignoud, M., Lederer, J., and Wainwright, M. J. (2016). A Practical Scheme and Fast Algorithm to Tune the Lasso With Optimality Guarantees. *Journal of Machine Learning Research*, 17(231):1–20.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, 9(3):432–441.

Meinshausen, N. and Bühlmann, P. (2006). Variable selection and High-dimensional graphs and with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Qin, L.-X., Zou, J., Shi, J., Lee, A., Mihailovic, A., Farazi, T. A., Tuschl, T., and Singer, S. (2020). Statistical Assessment of Depth Normalization for Small RNA Sequencing. *JCO Clinical Cancer Informatics*, 4:567–582.

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026.

Wang, T., Ren, Z., Ding, Y., Fang, Z., Sun, Z., MacDonald, M. L., Sweet, R. A., Wang, J., and Chen, W. (2016). FastGGM: An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLOS Computational Biology*, 12(2):e1004755.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of machine learning research : JMLR*, 13:1059–1062.