Supplementary Materials

*SIDEseq: a cell similarity measure defined by **s**hared **i**dentified **d**ifferentially **e**xpressed genes for single-cell RNA-sequencing data.*

Cell Culture

The human epithelial ovarian cancer cell line, CAOV-3, was obtained from American Type Culture Collection (ATCC, Manassas, VA, USA). Cells were maintained with Dulbecco's Modified Eagle's Medium (DMEM, Invitrogen, Carlsbad, CA, USA), supplemented with 10% fetal bovine serum (ATCC) and 1% penicillin and streptomycin (Invitrogen). Cell cultures were maintained at 37°C in the presence of 5% CO2, and routinely passaged once cultures reached 80% confluency. CAOV-3 cells were plated in 100 mm tissue culture dishes at a sub-cultivation ratio of 1:5, incubated overnight in supplemented DMEM medium, and then incubated with either thrombin (2.0 U/mL) or TGFβ-1 (5ng/mL) (both from R&D Systems, Minneapolis, MN, USA) for 48 hours. After treating with TGFβ-1 or thrombin, we used the DMEM (growth medium) and trypsin, to suspend the cells in growth medium.

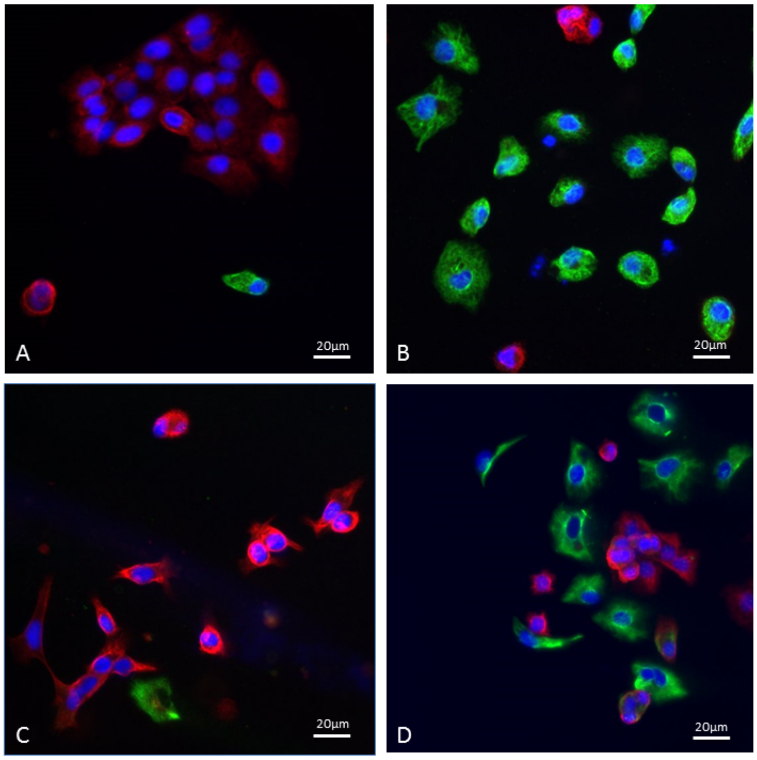


Fig 1. CAOV-3 were incubated for 48hrs in **A)** TGFβ-1 (0ng/mL), **B)** TGFβ-1 (5ng/mL), **C)** thrombin (0U/mL), and **D)** thrombin (2U/mL). The nucleus (blue, DAPI), vimentin (green, Alexa Fluor 488), and CK19 (red, Alexa Fluor 546) were stained.

Fluidigm C1 Preparation

Treated and untreated CAOV-3 were released from culture plates with 0.25% trypsin-EDTA (Corning, New York), rinsed with media, centrifuged at 3000 rpm, and re-supsended in media at a concentration of 100,000 cells/mL. The samples were then prepared by following the C1 single-cell auto prep system protocol outlined by Fluidigm before loaded onto the C1 System (Fluidigm, San Francisco, CA, USA). To prepare the sequencing-ready library for the Bioanalyzer QC and qPRCR step, a Nextera XT DNA Sample Preparation Kit was utilized.

Read Alignment and Quantification

We used three large scale datasets as a basis of analysis. To ensure a level of uniformity of the gene expression, we ran RNA short-reads from each experiment through the standard ENCODE pipeline using STAR [dobin2013star] for alignment and RSEM [li2011rsem] to call differential expression [ENCODE_pipe]. Brief descriptions of the experiments and data details can be found below.

Dobin, Alexander and Davis, Carrie A and Schlesinger, Felix and Drenkow, Jorg and Zaleski, Chris and Jha, Sonali and Batut, P..., "STAR: ultrafast universal RNA-seq aligner", *Bioinformatics* 29, 1 (2013), pp. 15--21.

Li, Bo and Dewey, Colin N, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome", *BMC bioinformatics* 12, 1 (2011), pp. 1.

ENCODE Consortium, Stadford University, UC Santa Cruz, "RNA-seq pipeline for long RNAs" (2016).

Normalization

There are data sets where the biological signals of interest are strong and overwhelm the unwanted variation in the data, in which case cells can cluster easily and successfully after simple normalization or even without normalization (Sandberg *et al*., 2014). However, this was not the case with the human ovarian cancer cell data. An important source of unwanted biological noise in scRNA-seq experiments, especially pertinent to our human ovarian cancer cell data set, is the variability introduced when cells to be sequenced have different passage numbers (ATCC, 2010). Passage number is defined as the number of times a cell culture was subcultured to maintain continued growth (ATCC, 2010). Passage number has a non-negligible effect on gene expression and regulatory pathways within cell lines (O'Driscoll *at al.,* 2006; Lin *et al.,* 2003). Thus, when cells are sequenced in different batches and the passage numbers are different, cells that were supposed to be biological replicates may become biologically different. This was the case for the human ovarian cancer cell data set, where a difference in passage number (i.e. differed by three) resulted in biologically different cells in the two batches. With a lack of "true" biological replicates between batches, this source of unwanted variation makes normalization across batches very challenging.

Indeed, exploratory data analysis reveals that the cells in the two batches vary significantly from each other. Plotting the cells according to their first two principal components through a PCA analysis shows a clear clustering of cells by batch (Fig. S1), especially along the first principal component, which explains roughly ten percent of the variance. A relative-log-expression plot of the TPM expressions also brings the difference between the two batches to the surface (Fig. S1). Of course, differences between the treated cells in the two batches is expected since the two inducers, TGFβ-1 and thrombin, are different. However, the untreated control cells in both batches should not have significantly different expression profiles, as seen in Figure S1. The apparent differences observed between the control cells in the two batches is likely due to technical noise or unwanted biological variability. Therefore, normalization of this data set is needed.

The human ovarian cancer cell data we analyzed faces a very complex normalization task since it contains various sources of wanted and unwanted variation (such as passage number effects). We tested three popular normalization methods with varying complexities. The first technique, full quantile normalization, is a simple non-linear normalization technique where the quantiles of the read count distributions over all samples are matched to a common reference distribution which is built using the median read counts across samples (Bullard *et al*., 2010). Full quantile normalization is more aggressive than global scaling normalization methods such as median normalization, which is likely necessary for this data set. We did full quantile normalization on raw read counts that were transformed to TPM values. The other two normalization techniques tested on the data set are remove-unwanted-variation techniques, named 'RUVg' and 'RUVs' from the 'RUVSeq' package (Risso *et al.,* 2014). The 'RUVSeq' package provides a few functions to remove unwanted factors of variation from RNA-seq data by using control genes or replicate samples, which are independent of the biological variability of interest, to estimate the factors of unwanted variation using factor analysis. In order to do the remove-unwanted-variation normalizations, we used raw read counts, as opposed to TPM or RPKM expression values, following the package instructions. The 'RUVg' method uses control genes to estimate the hidden factors of unwanted variation in data. The 'RUVs' method relies on both replicate samples and control genes to remove the unwanted factors of variation.

When performing 'RUVg' and 'RUVs' normalization, we used a list of human housekeeping gene from Eisenberg *et al.* (2013) as negative control genes. When 'RUVs' normalization was done across batches, we combined half of the untreated cells in each batch as replicate samples. When the normalization was done within batch we used half of the untreated cells within each. We believe that the untreated cells can be thought of as replicate samples because the factors of wanted variation (e.g. treatment with EMT inducing factors) are constant across these cells. However, using untreated cells as control samples requires that the identity of some of the cells be known, turning the normalization task into a semi-supervised problem. Furthermore, if normalization is to be done across batches, the untreated cells across batches must be viable control samples, which may not be the case due to passage-number effects. We tried normalization both across batches and within batch in order to compare the effects of this choice on clustering results.

The main assumption behind the methods in the 'RUVSeq' package is that the expected values of the log gene expressions for n samples with J genes are linear combinations of p factors of wanted variation, k factors of unwanted variation and an offset (Risso *et al.*, 2014).

$$E\left[Y_{nxJ}\,|W,X,O\right] = W_{nxk,unwanted} \times \alpha + X_{nxp,wanted} \times \beta + O \qquad (1)$$

One method in the 'RUVSeq' package, the 'RUVg' function, uses control genes, which are assumed to have constant expression across cells, to estimate the factors of unwanted variation. In the scRNA-seq setting, the assumption is that the control genes have constant expression across all cells, regardless of treatment, so any real variability in control gene expression across cells is assumed to be due to unwanted factors of variation such as technical noise. The factors of unwanted variation and their corresponding parameters are estimated using singular value decomposition and subtracted from the log gene expressions to obtain normalized counts. The method we used to select k for the human ovarian cancer cell data set was to begin with k equal to one and to increase k, if necessary, until the batch effect was removed. We use principal component plots and RLE plots to identify remaining batch effect in the normalized log counts after each value of this parameter was tested. We also used K-means clustering to test the stability of the normalized log count clusters after each tested value of the parameter. As a result of the above exploratory analysis, we used $k = 2$ factors of unwanted variation.

The 'RUVs' function in the 'RUVSeq' package relies on both replicate samples and control genes to remove the unwanted factors of variation. One assumption of this method is that the factors of wanted variation are constant across replicate samples (negative control samples may also be used). Therefore, the factors of wanted variation should not be contributing to the variation in gene expression seen across replicate samples. The second main assumption of the 'RUVs' function is the same assumption found in equation (1) above. As with 'RUVg' normalization, the factors of unwanted variation and their corresponding parameters are estimated and subtracted from the log gene expressions to obtain normalized counts. We used $k = 2$ factors of unwanted variation for the 'RUVs' method.

After ‘RUVg’ normalization of the ovarian cancer cell data across batches, a scatter plot of the first two principal components showed that cells no longer clustered by batch, but they also failed to cluster by treatment status (Supplementary Fig. S2). It is likely that ‘RUVg’ normalization removed too much variation, in this instance, when batches were normalized together. As mentioned previously, the different passage numbers of the batches likely introduced a large amount of unexpected, biological variation across batches, which the ‘RUVg’ method would determine to be unwanted technical noise (O’Driscoll *et al.,* 2006). This can be seen by the large proportion of variance explained by the first principal component alone in the combined-batch data, which diminishes when principal component analysis was done within batch (Fig. S1). After ‘RUVg’ normalization across batches, we performed hierarchical clustering using Pearson and Spearman correlation, resulting in no apparent clustering of cells by batch or by treatment status (Fig. S3). This agrees with the principal component plot and the concern that ‘RUVg’ normalization might have removed too much variation when used to normalize across batches. However, when we used the SIDEseq measure for clustering, it was interesting to see that cells still clustered almost perfectly by batch, even after ‘RUVg’ normalization across batches (Fig. S3). This suggests that the SIDEseq measure was able to explore the data at a deeper level and brought remaining, subtle differences between batches to the surface.

In all further analysis, we normalized and clustered cells in each batch separately. When ‘RUVg’ normalization was done within the TGFβ-1 group there was an improvement in the clustering of cells by treatment status in a principal component plot (Fig. S4). A relative log expression plot of the normalized data for the TGFβ-1 group showed a significant improvement in expression profiles (Fig. S4). However, when cells in the thrombin group were ‘RUVg’ normalized, they failed to cluster by treatment status (Fig. S6). These preliminary results suggest that the thrombin treatment may have been less effective than the TGFβ-1 treatment in inducing EMT, as it is harder to differentiate the thrombin cells from the untreated cells. Furthermore, this observation is supported by immunostaining images, which show that the thrombin treated cells have a smaller proportion of cells that have transitioned (see Supplementary Materials, Cell Culture). After ‘RUVg’ normalization within batch, hierarchical clustering using Spearman correlation of the cells in the TGFβ-1 group identified a small cluster of control cells and a cluster of mostly treatment cells (Fig. S5), but most cells did not cluster well by treatment status. When the SIDEseq measure was used, there was no improvement of the clustering of cells by treatment status (Fig. S5). For both Spearman correlation and the SIDEseq measure, the ability of cells to cluster by treatment status diminished with the cells from the thrombin group.

After full quantile normalization of log TPM expressions, the TGFβ-1 group showed little improvement in clustering results (Fig. S7). A principal component plots showed that cells in the thrombin group do not cluster by treatment status (Fig. S9). When hierarchical clustering was done within batch using Spearman correlation and the SIDEseq measure, there was no improvement over ‘RUVg’ normalization (Fig. S8). Next, ‘RUVs’ normalization and clustering were performed. This normalization method that requires replicate samples should not be done across batches for the human ovarian cancer cell data because the untreated cells in TGFβ-1 group and the thrombin group likely became biologically different as a result of different passage numbers, and can therefore no longer be thought of as control samples. A principal component plot showed an improvement in the separation of the TGFβ-1 group cells by treatment status after ‘RUVs’ normalization (Fig. S10). A principal component plot of the cells in the thrombin group after ‘RUVs’ normalization showed no improvement in the separation of cells by treatment status (Fig. S12). Therefore, all three normalization methods failed to improve the clustering of cells in the thrombin group, providing convincing evidence that the thrombin treatment cells did not differentiate significantly from the untreated cells. This observation concerning the thrombin treatment is an important biological observation from this data set and merits further investigation.

The ‘RUVs’ normalized counts in batch 1 showed the most promise when clustering cells by treatment and control status. We clustered the ‘RUVs’ normalized counts in the TGFβ-1 group using hierarchical clustering with Euclidean distance, Pearson and Spearman correlation, and the SIDEseq similarity measure. When looking at the dendrogram corresponding to the SIDEseq measure, three clusters were recognized. One was a large cluster consisting of only untreated cells. Actually most untreated cells were in this cluster. Another cluster consisted of a mix of treated and untreated cells. The third was a large cluster of mostly all treated cells. This cluster was on the outside of the sub-dendrogram formed by the other two clusters. In addition to clearer clusters of cells, the organization of the clusters within the dendrogram is also biologically interesting. The cluster that contained a mix of treated and untreated cells may correspond to a group of cells in the beginning stages of EMT or that have not entirely transitioned to the mesenchymal phenotype. The treated cells within the mixed cluster would then be more biologically similar to the untreated cells.

Flow chart showing the creation of the SIDEseq dissimilarity matrix:

STEP 1

N cells

J genes

$$\begin{bmatrix} x_1^1 & \cdots & x_N^1 \\ \vdots & \ddots & \vdots \\ x_1^J & \cdots & x_N^J \end{bmatrix}_{J\times N} \quad x_j^k = \text{expression of gene } k \text{ in cell } j$$

STEP 2

$$\begin{bmatrix} T_{12}^1 & \cdots & T_{1N}^1 \\ \vdots & \ddots & \vdots \\ T_{12}^J & \cdots & T_{1N}^J \end{bmatrix}_{J\times(N-1)} ,\cdots, \begin{bmatrix} T_{21}^1 & \cdots & T_{2N}^1 \\ \vdots & \ddots & \vdots \\ T_{21}^J & \cdots & T_{2N}^J \end{bmatrix}_{J\times(N-1)} ,\cdots, \begin{bmatrix} T_{N1}^1 & \cdots & T_{N(N-1)}^1 \\ \vdots & \ddots & \vdots \\ T_{N1}^J & \cdots & T_{N(N-1)}^J \end{bmatrix}_{J\times(N-1)}$$

$$T_{i,j}^k = \text{differential expression statistic between cell } i \text{ and cell } j \text{ for gene } k$$

STEP 3

DE Matrix 1 DE Matrix 2 DE Matrix N

$$\begin{bmatrix} G_{11}^{(1)} & \cdots & G_{1N}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{11}^{(n)} & \cdots & G_{1N}^{(n)} \end{bmatrix}_{n\times(N-1)} , \begin{bmatrix} G_{21}^{(1)} & \cdots & G_{2N}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{21}^{(n)} & \cdots & G_{2N}^{(n)} \end{bmatrix}_{n\times(N-1)} ,\cdots, \begin{bmatrix} G_{N1}^{(1)} & \cdots & G_{N(N-1)}^{(1)} \\ \vdots & \ddots & \vdots \\ G_{N1}^{(n)} & \cdots & G_{N(N-1)}^{(n)} \end{bmatrix}_{n\times(N-1)}$$

$$DE\ Matrix\ i = \text{differential expression matrix for cell } i.$$

$$G_{i,j}^{(k)} = \text{name of gene with the } kth \text{ largest DE statistic between cell } i \text{ and cell } j \ (i.e.\ T_{i,j}^{(k)}).$$

STEP 4

$$S_{i,j} = \sum\nolimits_{t=1,\dots,N,t\neq i,j} \left| \left(G_{it}^{(1)},\dots,G_{it}^{(n)}\right) \cap \left(G_{jt}^{(1)},\dots,G_{jt}^{(n)}\right) \right| \text{ similarity measure between cell } i \text{ and cell } j$$

STEP 5

$$S = \begin{bmatrix} 0 & \cdots & S_{N1} \\ \vdots & \ddots & \vdots \\ S_{1N} & \cdots & 0 \end{bmatrix}_{N\times N} \quad \text{The SIDEseq similarity matrix}$$

STEP 6

$$D = \begin{bmatrix} 0 & \cdots & Max - S_{N1} \\ \vdots & \ddots & \vdots \\ Max - S_{1N} & \cdots & 0 \end{bmatrix}_{N\times N} \quad \text{The SIDEseq dissimilarity matrix (Max=maximum in } S\text{)}$$

An alternative to step four is to divide the numbers of genes in the intersection by the number of genes in the union. This alternative similarity measure is related monotonically to the original measure used in SIDEseq. Clustering results using this alternative measure are nearly equivalent to those using the original measure proposed in step four.

$$S_{i,j} = \sum_{t=1,\ldots,N,t\neq i,j} \frac{\left|\left(G_{it}^{(1)},\ldots,G_{it}^{(n)}\right) \cap \left(G_{jt}^{(1)},\ldots,G_{jt}^{(n)}\right)\right|}{\left|\left(G_{it}^{(1)},\ldots,G_{it}^{(n)}\right) \cup \left(G_{jt}^{(1)},\ldots,G_{jt}^{(n)}\right)\right|}$$

Toy Example Explaining SIDEseq

Below is a toy example showing some of the benefits of the SIDEseq similarity measure, and how it is able to bypass noise in the expression levels of cells to get at true subpopulations. Gene 1 and gene 2 are differentially expressed between the two subpopulations of the example, while the other genes are uninformative and simply provide noise. SIDEseq is able to identify the subpopulations, even though there is noise, why the other measures cannot.

(A) Toy Data

		gene 1	gene 2	gene 3	gene 4	gene 5	gene 6	gene 7	gene 8	gene 9	gene 10	gene 11
Subpopulation 1	cell A	5.5	5.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5
	cell B	5.5	5.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0
Subpopulation 2	cell C	4.5	4.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5
	cell D	4.5	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0	4.5	4.0

(B) Some commonly used (dis)similarity measures and the SIDEseq measure computed based on the above data

Pearson Correlation

	cell B	cell C	cell D
cell A	0.64	**0.72**	0
cell B		-0.06	**0.77**
cell C			-0.69

Spearman Correlation

	cell B	cell C	cell D
cell A	0.05	**0.9**	-0.31
cell B		-0.39	**0.93**
cell C			-0.69

Euclidian Distance

	cell B	cell C	cell D
cell A	1.5	**1.41**	2.06
cell B		2.06	**1.41**
cell C			1.5

SIDEseq (top 2 DE genes used to define SIDEseq)

	cell B	cell C	cell D
cell A	**2**	0	0
cell B		0	0
cell C			**2**

SIDEseq (top 3 DE genes used to define SIDEseq)

	cell B	cell C	cell D
cell A	**>=1**	**<=0.4***	<=0.4
cell B		<=0.4	<=0.4
cell C			**>=1**

***SIDEseq for cell A and cell C when top 3 DE genes are considered.** The computation is based on comparing the DE matrices for cell A and cell C in **(C)**. Looking at the DE gene lists under "A vs. B" and "C vs. B" (cells A and C are separately compared with cell B), the interaction of top 3 in the two lists will have either 0 genes (then 6 genes in the union) or 1 gene (then 5 in the union). There will be similar results when comparing the DE lists under "A vs. D" and " C vs. D". Thus the SIDEseq measure will be at best =1 /5 + 1/5 = 0.4 between cell A and cell C.

(C) Matrices of ordered DE genes between cells

DE matrix for cell A

	A vs. B	A vs. C	A vs. D
Genes ordered by T	genes 3-11 (T=0.16)	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)
	genes 1-2 (T= 0)	genes 3-11 (T=0)	genes 3-11 (T=0.16)

T = |x-y|/sqrt(x+y): differential expression statistic

DE matrix for cell B

	B vs. A	B vs. C	B vs. D
Genes ordered by T	genes 3-11 (T=0.16)	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)
	genes 1-2 (T= 0)	genes 3-11 (T=0.16)	genes 3-11 (T=0)

DE matrix for cell C

	C vs. A	C vs. B	C vs. D
Genes ordered by T	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)	genes 3-11 (T=0.16)
	genes 3-11 (T=0)	genes 3-11 (T=0.16)	genes 1-2 (T= 0)

DE matrix for cell D

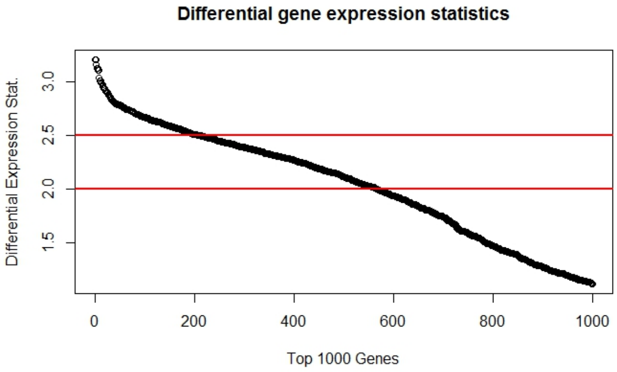
	D vs. A	D vs. B	D vs. C
Genes ordered by T	genes 1-2 (T=0.32)	genes 1-2 (T=0.32)	genes 3-11 (T=0.16)
	genes 3-11 (T=0.16)	genes 3-11 (T=0.)	genes 1-2 (T= 0)

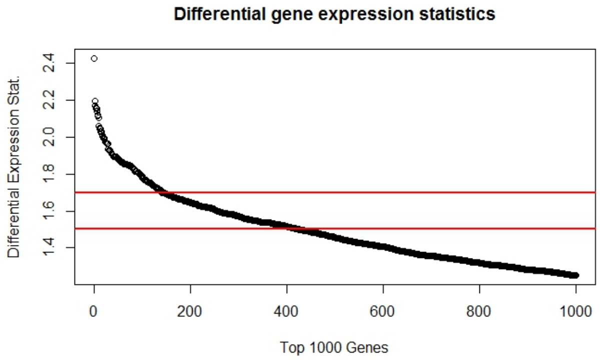
Exploration and Validation of the SIDEseq Measure

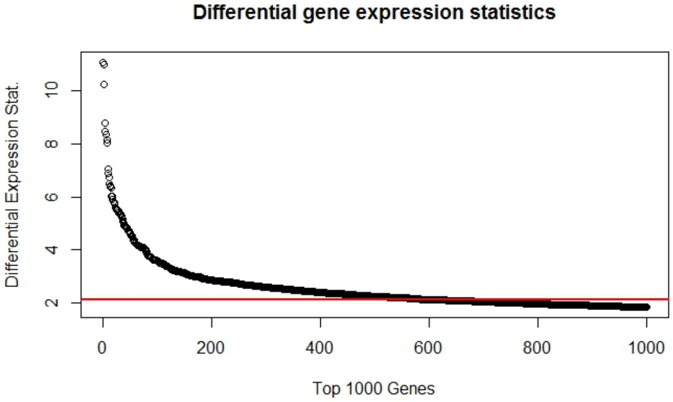
NOTE: To perform hierarchical clustering using Euclidean distance, Pearson correlation, Spearman correlation or using the SIDEseq measure, we used the 'hclust' function in the 'stats' package in R (version 3.2.5), specifying the 'ward.D2' method. We found that the 'ward.D2' method generally resulted in clearer clusters for all distance measures when used on the scRNA-seq data sets, as opposed to the default 'complete linkage' method.

Selecting the number of differentially expressed genes (n)

To select the parameter n, the cutoff for the number of differentially expressed genes, for the novel distance measure, plots such as those shown below were used. For a given data set, we chose n by plotting the differential expression statistics between various cells, and choosing a range of values for n which captured enough information without including too many noisy genes. The three plots below correspond to differential statistics between two cells in the human cancer cell data set from Ramsköld *et al*. (2012) (top left), in the human embryo cell data set from Yan *et al.* (2012) (top right), and in the human ovarian cancer cell data set after 'RUVs' normalization (bottom).







Considering different values of (n) for different subpopulations

Should SIDEseq allow the value of n to change between subpopulations? For example, consider the case where there are three subpopulations S1, S2, S3, and that there are n1 DE genes between S1 and S2 and n2 DE genes between S2 and S3. If n1 is a lot smaller than n2, how does this affect the performance of SIDEseq? Let's consider this example where n1 << n << n2. Since n1 << n, a list of size n of DE-genes between S1 and S2 cells would be very noisy (containing a lot of non-DE, random genes), especially when it is compared with the lists derived between S3 and S2. However, we want to point out that *no matter the lists are noisy or not, as long as there is a reasonable mix of noisy and informative DE genes lists (this usually can be achieved with n not too far from the median), all the lists together will provide useful information to help distinguish cells from S1, S2 and S3*. To see this point, let us further assume that there are n3 true DE genes between S1 and S3. Without loss of generality, let us assume n3 is also very small, that is, both n1 and n3 are small. Arguments will be similar when n3 is large or of reasonable size.

In the above situation, we expect:

- The cells from S1 would have the property that all their associated DE gene lists are noisy. Then the SIDEseq values between an S1 cell and any other cell would be small (since it is unlikely to observe a significant overlap between DE gene lists if at least one list has been noisy).
- The cells from S2 would have the property that they have informative DE gene lists against S3 cells but noisy DE gene lists against S1 cells or S2 cells. Then the SIDEseq values between two S2 cells would be reasonably large since their associated informative DE gene lists against S3 cells would significantly overlap. However, the SIDEseq values between an S2 cell and an S3 cell would be small since their associated informative DE gene lists are always against different cells (i.e., when compared with S2 cells, S3 cells will have informative DE gene lists while S2 cells will have noisy DE gene lists; When compared with S3 cells, S2 cells will have informative DE gene lists while S3 cells will have noisy DE gene lists).
- The properties of S3 cells can be similarly argued as above. In brief, the SIDEseq values between two S3 cells would be reasonably large.

In summary, S2 and S3 subpopulations can be well identified. S1 cells show different properties from S2 and S3 cells but it is hard to claim that the S1 cells form their own cluster since they have small SIDEseq values among themselves. This however does not seem that unreasonable to us since exceptionally small n1 and n3 may mean that the subpopulation S1 does not possess "convincing characteristics" to form its own subpopulation. Moreover, S2 cells and S3 cells have stronger subpopulation-specific functional homogeneity compared to the S1 cells and thus the smaller SIDEseq measures between S1 cells do seem to make sense to us. Of course this point is debatable. We also note that as n1 and n3 increase, the SIDEseq values between cells within S1 would increase too. Also if there is another subpopulation under consideration, and there is a long list of DE genes between S1 and this subpopulation, then the SIDEseq values between cells within S1 would become large (that is, SIDEseq is informative to identify S1 now).

If the numbers of true DE genes between all pairs of two cells are available, we might prefer to choose "*n*" close to the median. However, in practice, we would not know what the actual numbers of DE genes are. So we usually choose *n* from 150-1000, since that is the size of most lists of biologically meaningful DE genes people have reported in different studies.

Although we have been largely happy with using the DE gene lists of the same length, we admit that SIDEseq would be further improved if we could effectively take into account the varied lengths of different DE gene lists. We consider achieving this by adapting the idea in Li et al (2011)[1]. The work in that paper concerns the reproducibility of findings (e.g., identified peaks from a ChIP-sep experiment) from replicate experiments. Particularly, that paper defines reproducibility as the extent to which the ranks of the significance of the findings are consistent across replicates. By jointly modeling the ranks of findings from replicate experiments using a copula mixture model, a score called the irreproducible discovery rate (IDR), analogous to a false discovery rate, was derived to measure reproducibility.

To apply IDR to our analysis, for any two cells, we will first compute our differential statistics, T's, for all genes. That is, we can obtain a T-profile for each pair of two cells. In the context of our study, we are interested in comparing two T-profiles: genes associated with high T-values are likely DE genes; if there are a lot of common genes with high T-values in both profiles, the two profiles then share a lot of common DE genes, and also a stronger dependence among high T-values in the two profiles. Accordingly, we now consider that the bivariate data ($X_1,\ldots, X_n, Y_1,\ldots, Y_n$) from two T-profiles consisting of genuine signals (i.e., overlapping DE genes or positively correlated high X_i's and Y_i's) and spurious signals (i.e., non-overlapping DE genes and other random genes or uncorrelated X_i's and Y_i's). Let π_1 and $\pi_0 = 1 - \pi_1$ denote the proportion of overlapping DE genes (Z_i=1) and the rest of the genes (Z_i=0), respectively. We further assume X_i and Y_i are from a continuous bivariate distribution with density h_1 given Z_i=1 (respectively, h_0 given Z_i=0). The mixture copula model can then be expressed as

$$\varphi(T_1(x), T_2(y), \theta_0, \theta_1) = \{\pi_0 h_0(T_1(x), T_2(y), \theta_0) + \pi_1 h_1(T_1(x), T_2(y), \theta_1)\} T'_1(x) T'_2(y),$$

with h_1 and h_0 describing different dependence levels between X and Y. $T_1(x_i)$ and $T_2(Y_i)$ are the unknown scales, which can be estimated empirically. After fitting the copula mixture model, based on the estimated π_1 and π_0 and the two fitted distributions h_1 and h_0, we can then estimate the chance that a gene is a common DE gene between the two lists (i.e., $P(Z_i = 1 | X_i, Y_i)$) by

$$z(g_x, g_y, i) = \frac{\hat{\pi}_1 \hat{h}_1(\hat{T}_1(x_i), \hat{T}_2(y_i))}{\hat{\pi}_0 \hat{h}_0(\hat{T}_1(x_i), \hat{T}_2(y_i)) + \hat{\pi}_1 \hat{h}_1(\hat{T}_1(x_i), \hat{T}_2(y_i))}.$$

There are two ways to use the estimated parameters from the above model.

- Since h_1 describes the "dependent" component between X and Y, we can use the estimated dependence parameter associated with h_1 directly to reflect the level of consistency in terms of DE genes between two T-profiles. A SIDEseq measure can then be defined accordingly by replacing the number of intersected DE genes, denoted by $S_{i,k}$ in the flow chart of the Supplementary Material, by the estimated dependence parameter.
- We can classify the genes based on the estimated $z(g_x, g_y, i)$ to obtain a set of common DE genes between two T-profiles. A SIDEseq measure can then be defined accordingly by replacing the number of intersected DE genes, denoted by $S_{i,k}$ in the flow chart of the Supplementary Material, by the size of the inferred set of common DE genes.

Since IDR has an installed R-package, a Gaussian copula version of the above method can be easily implemented. However, we note that this approach can be quite time consuming if there are many cells to study, in which case, the simple method based on DE gene lists with the same size would likely be more favorable.

Simulations

Splatter Simulations

To simulate larger scRNA-seq data sets for the GiniClust comparison, we used the R package ‘splatter’ from the Oshlack lab (https://github.com/Oshlack/splatter). We used the default parameters set by the authors, and then made different adjustments to the parameters for each simulation in order to make the clustering of cells into subpopulations more difficult. We changed the ‘groupCells’ parameter to increase the number of subpopulations, the ‘de.prob’ parameter to change the probability of differential expression for the genes in each subpopulation, and the ‘de.facLoc’ to change the degree of differential expression. In every simulation we specified the ‘method=”groups”’ argument in order to simulate data sets with subpopulations of cells. To get the data set where subpopulations have different probabilities of differential expression, we simulated 5 subpopulations separately, each time changing the differential expression probability, but setting the same seed so that the subpopulations could be combined. As described in the paper, the GiniClust algorithm was then used on each simulated data set in order to identify the top Gini index genes to use in a comparison with SIDEseq via hierarchical clustering, and to assess the overall performance of the GiniClust method.

In all simulations, SIDEseq performed better than the traditional similarity measures, as evidenced by the ARI values and visual inspection of the dendrograms (Table 1 and Figures S18 and S19). SIDEseq’s ability to very accurately identify subpopulations in settings where each subpopulation has a different probability of differential expression or when the factor of differential expression is lowered is very encouraging.

There are some promising results when the top Gini index genes were used with common similarity measures for hierarchical clustering. However, we were surprised at the poor performance of the GiniClust algorithm when it was run as a whole on the simulated data sets and on our human ovarian cancer cell data set. In all cases, all cells were grouped into the same cluster, while a few cells were identified as rare cell types. We believe this is because there were insufficient numbers of genes which passed the normalized Gini index threshold, and so not enough genes were available for clustering in the DBSCAN algorithm that GiniClust utilizes.

Small Simulations

In order to better understand the benefits of the SIDEseq measure for sequencing data, we simulated small scRNA-seq data sets and evaluated the ability of the SIDEseq measure, Euclidean distance and Pearson and Spearman correlation to accurately capture the relationships between cells. The simulated data sets consist of 1,000 gene expression for four cells, where two of the cells come from one subpopulation and the other two cells come from another. Each subpopulation is defined by a subset of 10 differentially expressed genes.

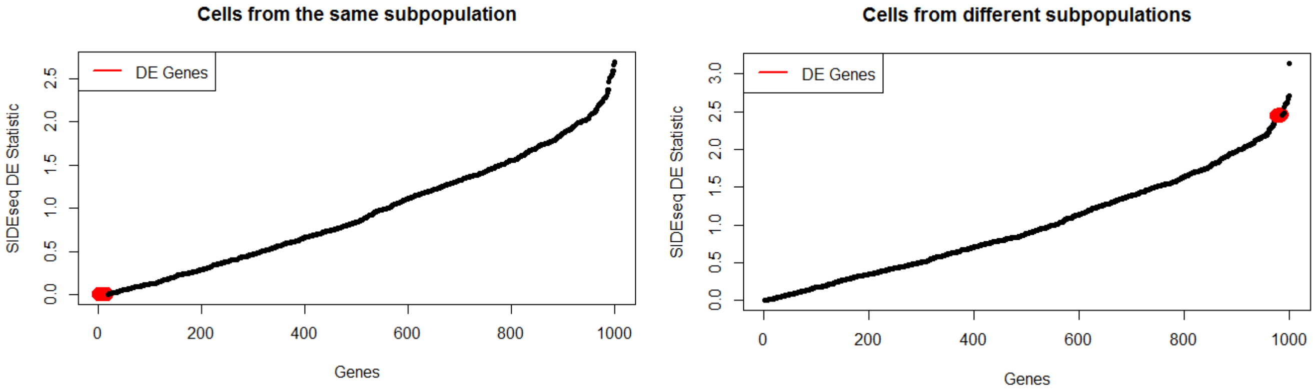
Subpopulation 1:

Cells A and B: 10 genes $\sim Normal(\mu = 6, \sigma^2 = 0.01^2)$, 990 genes $\sim Normal(\mu = 2, \sigma^2 = 1.7^2)$

Subpopulation 2:

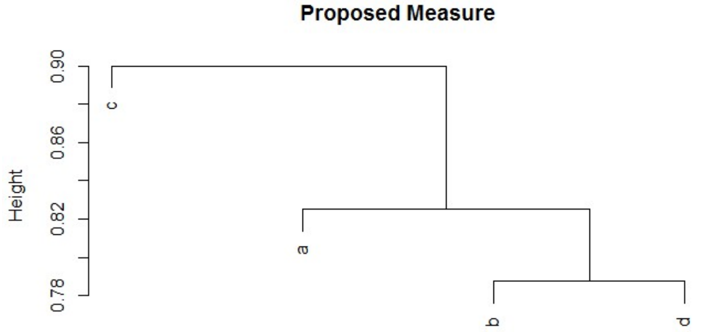
Cells C and D: 10 genes $\sim Normal(\mu = 0, \sigma^2 = 0.01^2)$, 990 genes $\sim Normal(\mu = 2, \sigma^2 = 1.7^2)$

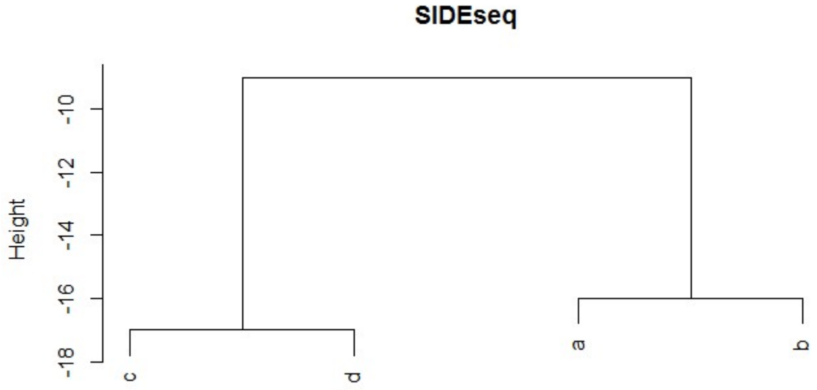
The above data set was generated several times to ensure the robustness of the results. Euclidean distance frequently failed to identify the correct relative similarities between cells and to cluster them correctly by subpopulation. This is because only a small set of genes are differentially expressed between the subpopulations (1%), and the variability in the expressions for the non-differentially expressed genes overwhelmed the difference in expressions for the differentially expressed genes. Pearson correlation (and similarly Spearman correlation) performed worse than Euclidean distance on this data set, for similar reasons. When the SIDEseq measure is used, however, it almost always correctly clusters the cells by subpopulation. The SIDEseq measure works because if two cells come from the same subpopulation, they will share many genes in common identified as top differentially expressed genes with the cells in the other subpopulation. A high level of consistency in the top genes identified as differentially expressed with other cells implies a high SIDEseq similarity measure. For example, it is likely that the ten genes which are differentially expressed between cells A and D will be among the top differentially expressed genes as identified by the differential expression statistic in the SIDEseq measure. Many of the genes which are in the top identified differentially expressed genes between cells A and D will also be in the top genes identified as differentially expressed between cells B and D, since cell D is in a different subpopulation. This causes the SIDEseq similarity measure between cells A and B to be high. Cells from different subpopulations, like cells A and C, will not share a lot of genes which are identified as differentially expressed with the other cells, and will therefore have low SIDEseq similarity measures.



Simulations, Positive and Negative Differential Expression:

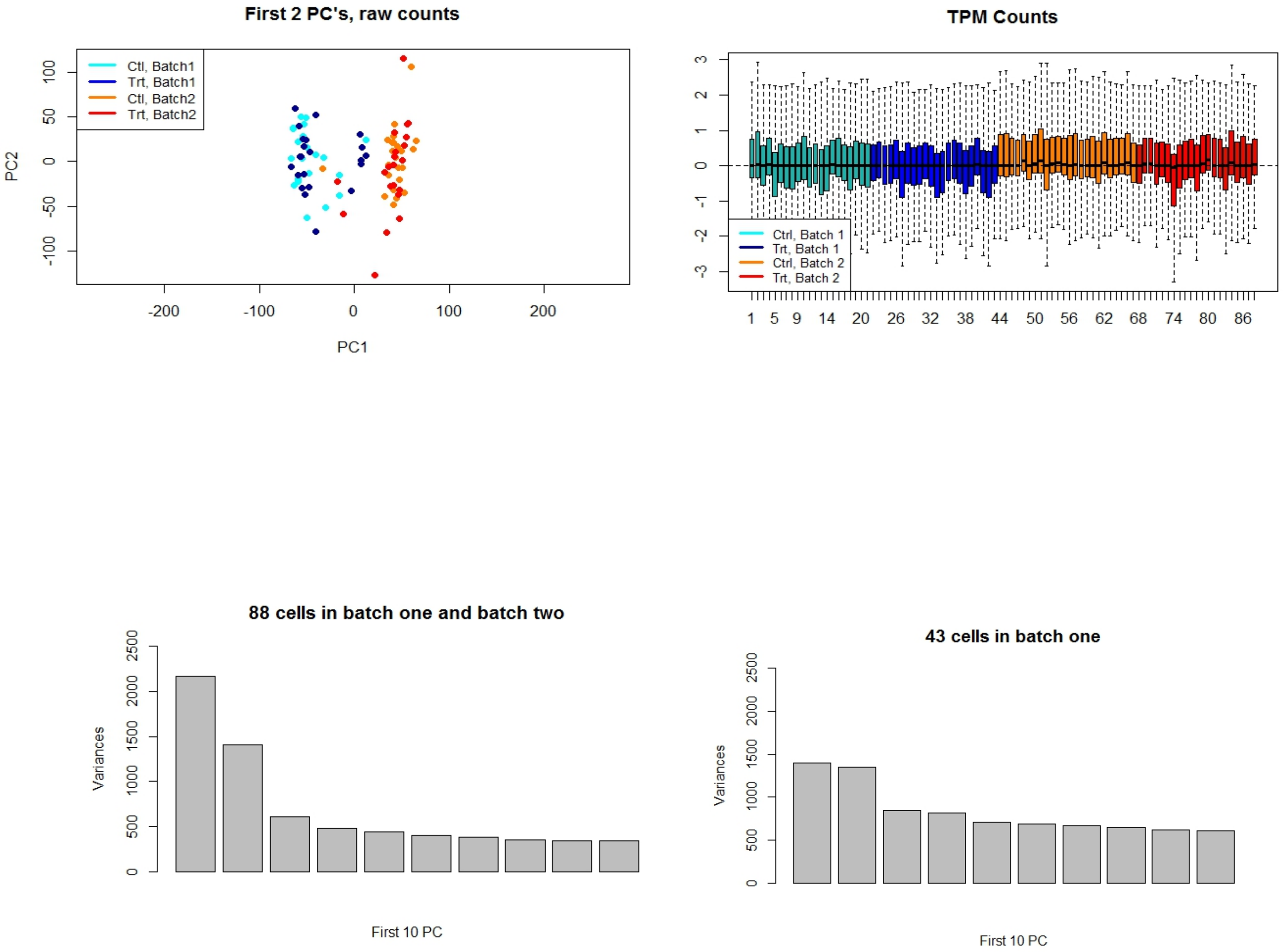
A suggestion for a similarity measure for scRNA-seq data, which is similar to SIDEseq but may improve upon it, could be the following: to find the similarity between cell i and cell j, separate cell i and j from the population of cells. Call this remaining group of cells which does not include cells i and j the "subpopulation". Find the mean expression level for each gene within the subpopulation. Next, identify the set of genes which are positively and negatively differentially expressed between cell i and the mean expressions of the subpopulation. Do the same for cell j. Then, the similarity between cell i and cell j is the number of genes which they have in common which are positively differentially expressed with the average expression of the subpopulation plus the number of negatively differentially expressed genes they have in common with the mean expressions of the subpopulation. This is similar to the SIDEseq measure with the exception of two big differences: differentially expressed genes are separated by direction of differential expression, and differentially expressed genes are found with respect to the *average* expression levels within the subpopulation. We used the same simulated data set above to compare the performance of SIDEseq with this proposed measure. The data set was generated repeatedly, each time calculating two dissimilarity matrices from each of the two measures and using them to perform hierarchical clustering. The SIDEseq measure outperformed the proposed measure each time, since it was able to group cells A and B together and cells C and D together. The proposed measure sometimes grouped cells A and B, but failed to group cells C and D or failed altogether to capture the correct relationships (see dendrogram below). These results suggest two ideas. One is that splitting up differentially expressed genes by sign does not necessarily improve the performance of similarity measures. The other idea for why SIDEseq performs better is because the proposed measure treats the remaining cells as one population and averages their expression levels. In highly heterogeneous data sets with a lot of variability, such as the simulation data set or the human ovarian cancer cell data set, this may cause the similarity measure to perform poorly.


Proposed Measure

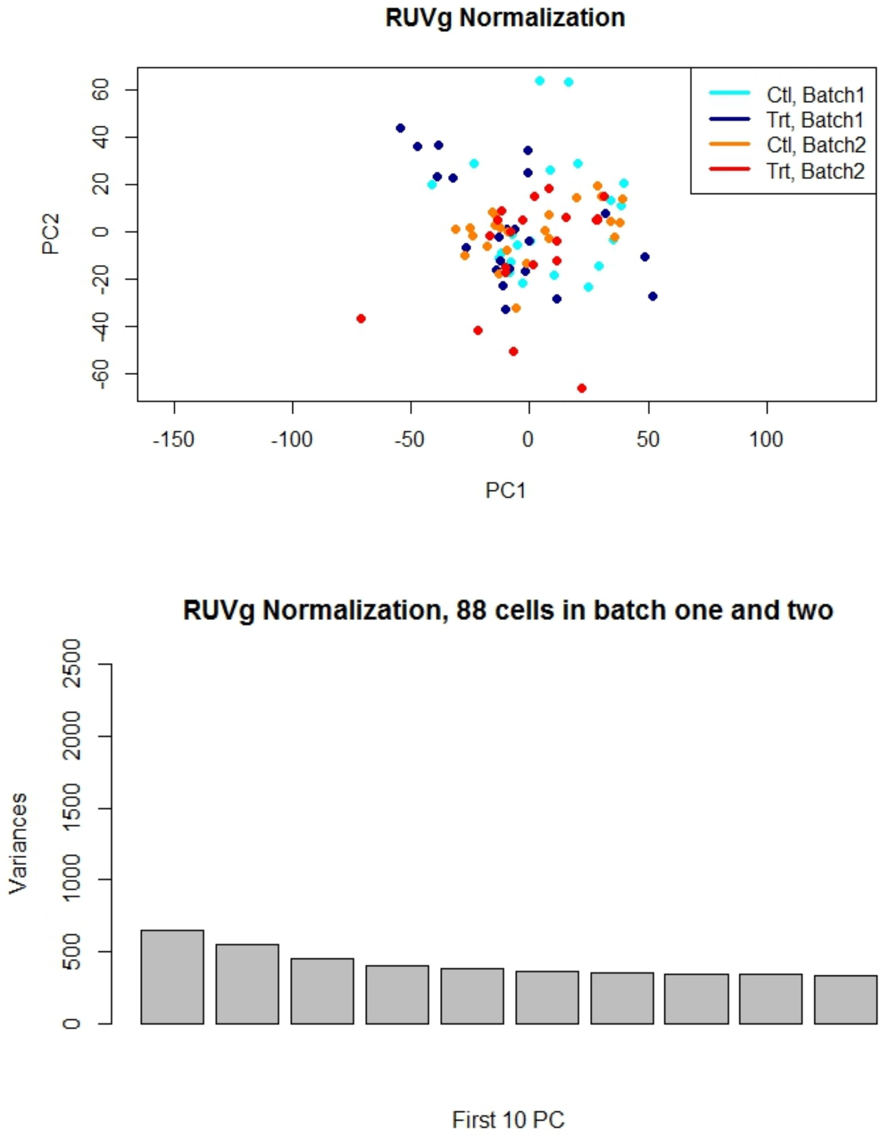

SIDEseq

Supplementary Figures

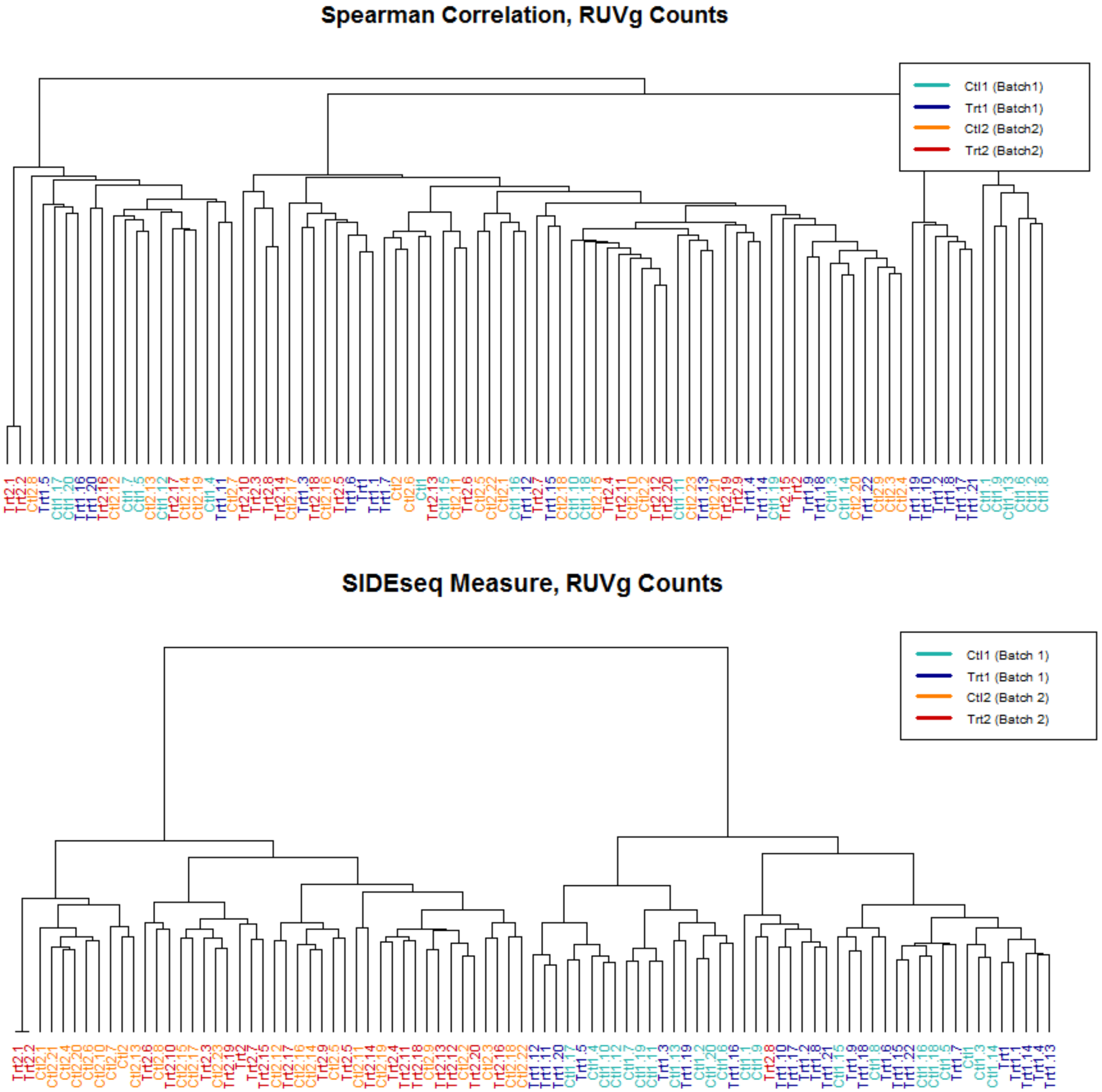
(In all supplementary figures, batch one refers to the TGFβ-1 group and batch two refers to the thrombin group.)



Supplementary Figure S1: Principal component analysis on the 88 human ovarian cancer cells from batches one and two reveals that roughly ten percent of the total variance is explained by the first principal component. When principal component analysis is done only within batch one, the large proportion of variance found in the first principal component diminishes. This suggests that there exists a significant amount of technical and unwanted biological variation across batches, and that perhaps normalization and clustering should be done within batch.

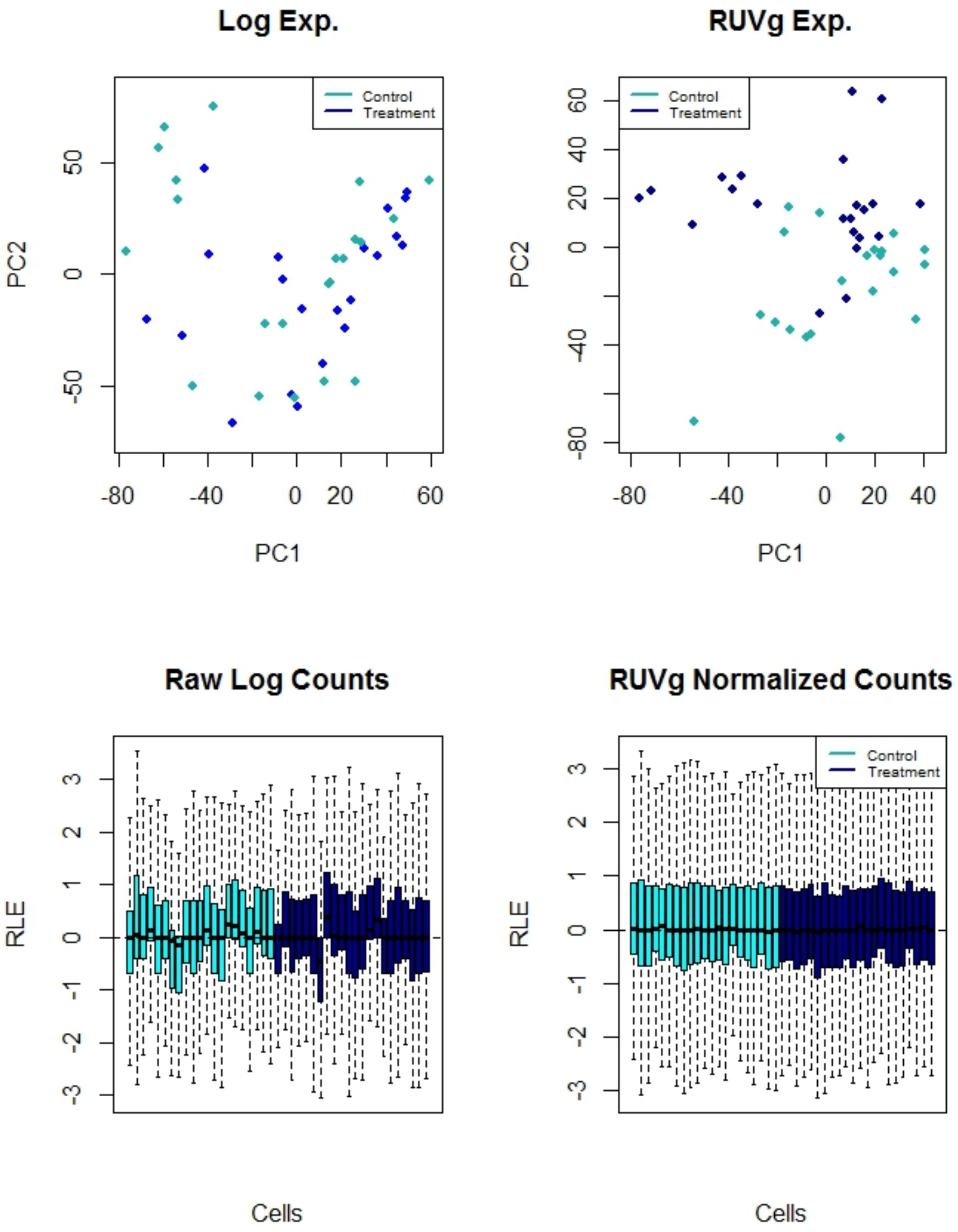


<u>**Supplementary Figure S2:**</u> After ‘RUVg’ normalization of the 88 human ovarian cancer cells in batch one and batch two (with $k = 2$ unwanted factors of variation), principal component analysis reveals that much of the variation within the cells has been removed. Cells no longer cluster by batch, but they also do not cluster by treatment and control status. The large proportion of variance explained by the first principal component has diminished. ‘RUVg’ normalization may have removed too much information when cells from both batches were normalized together, suggesting that normalization should be done within batch. This does not change even when $k = 1$ is specified for the normalization.

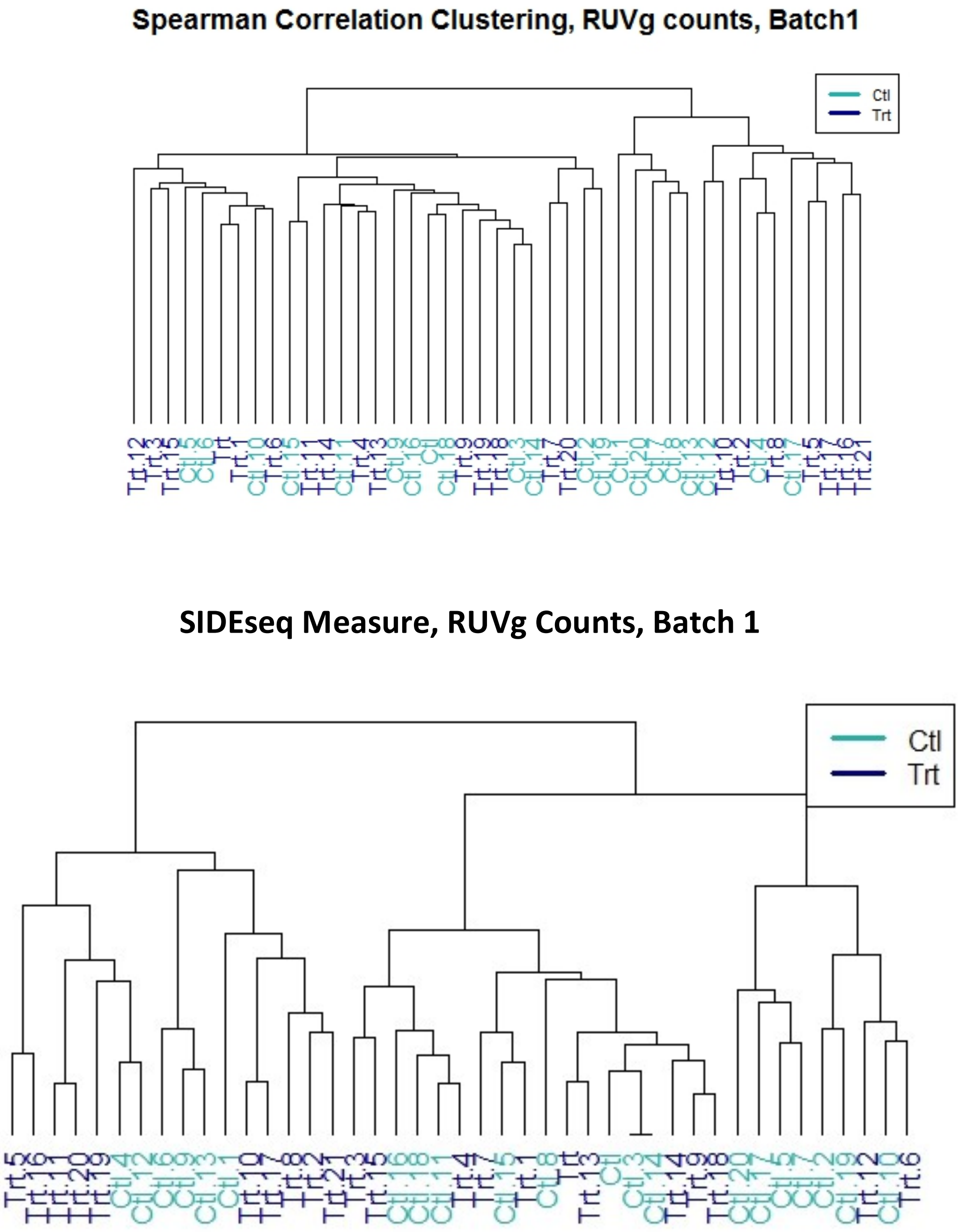

Spearman Correlation, RUVg Counts

SIDEseq Measure, RUVg Counts

<u>Supplementary Figure S3:</u> When Spearman correlation is used to do hierarchical clustering of cells together in both batches, the result is consistent with the principal component plot of Supplementary Figure S2. Cells do not cluster by batch, nor do they cluster by their treatment and control status. However, when the SIDEseq measure is used, cells cluster almost perfectly by batch. This suggests that the SIDEseq measure was able to uncover the remaining differences between the batches, even after normalization, and even when a principal component plot and hierarchical clustering using Spearman correlation did not. Results using Pearson instead correlation are similar.

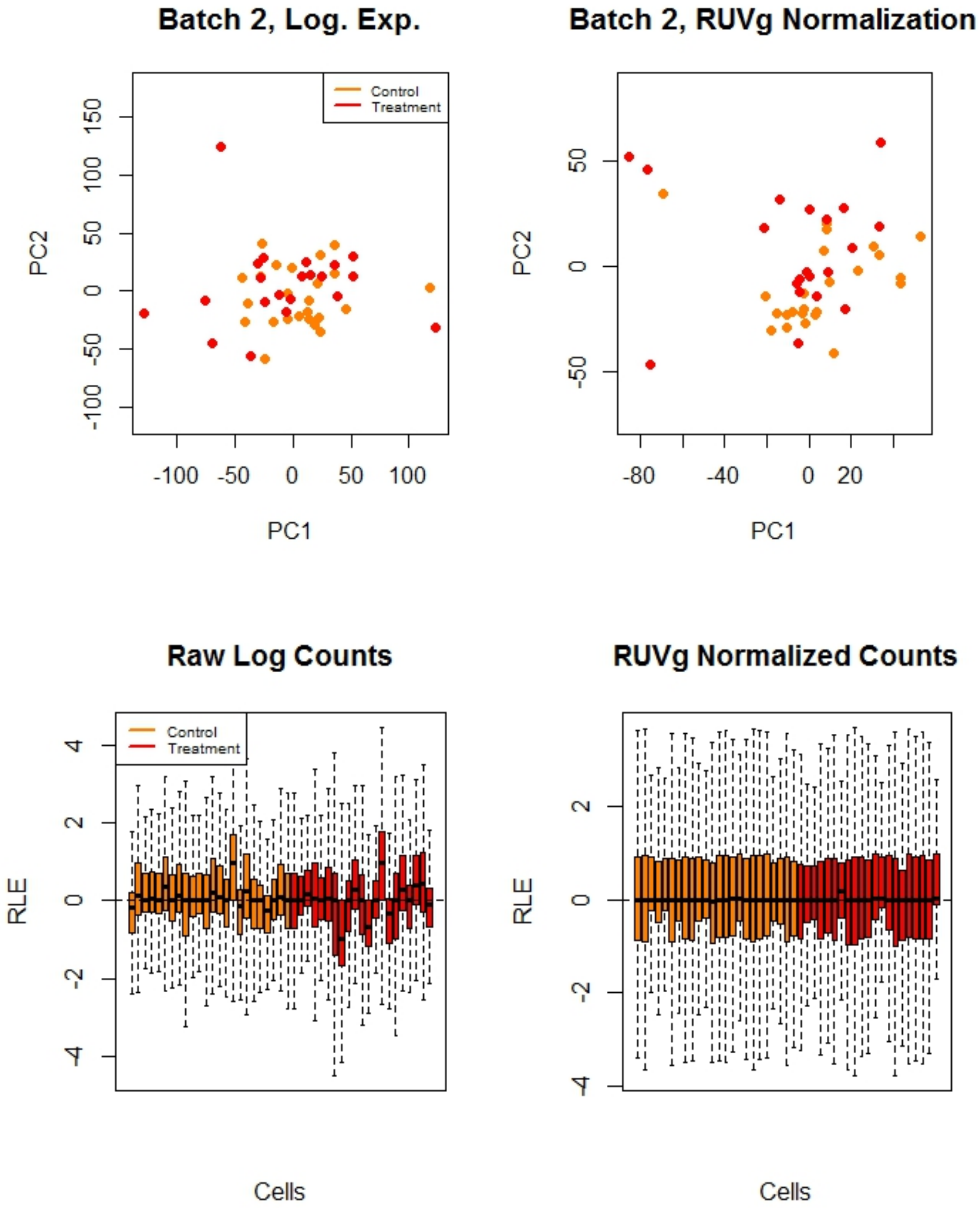


<u>Supplementary Figure S4:</u> 'RUVg' normalization of the 43 human ovarian cancer cells in batch one shows both an improvement in the separation of cells by treatment and control status, as well as in the expression profiles of the cells. Un-normalized log expressions of cells in batch one were also used to plot the cells according to their first and second principal component for comparison (top left).

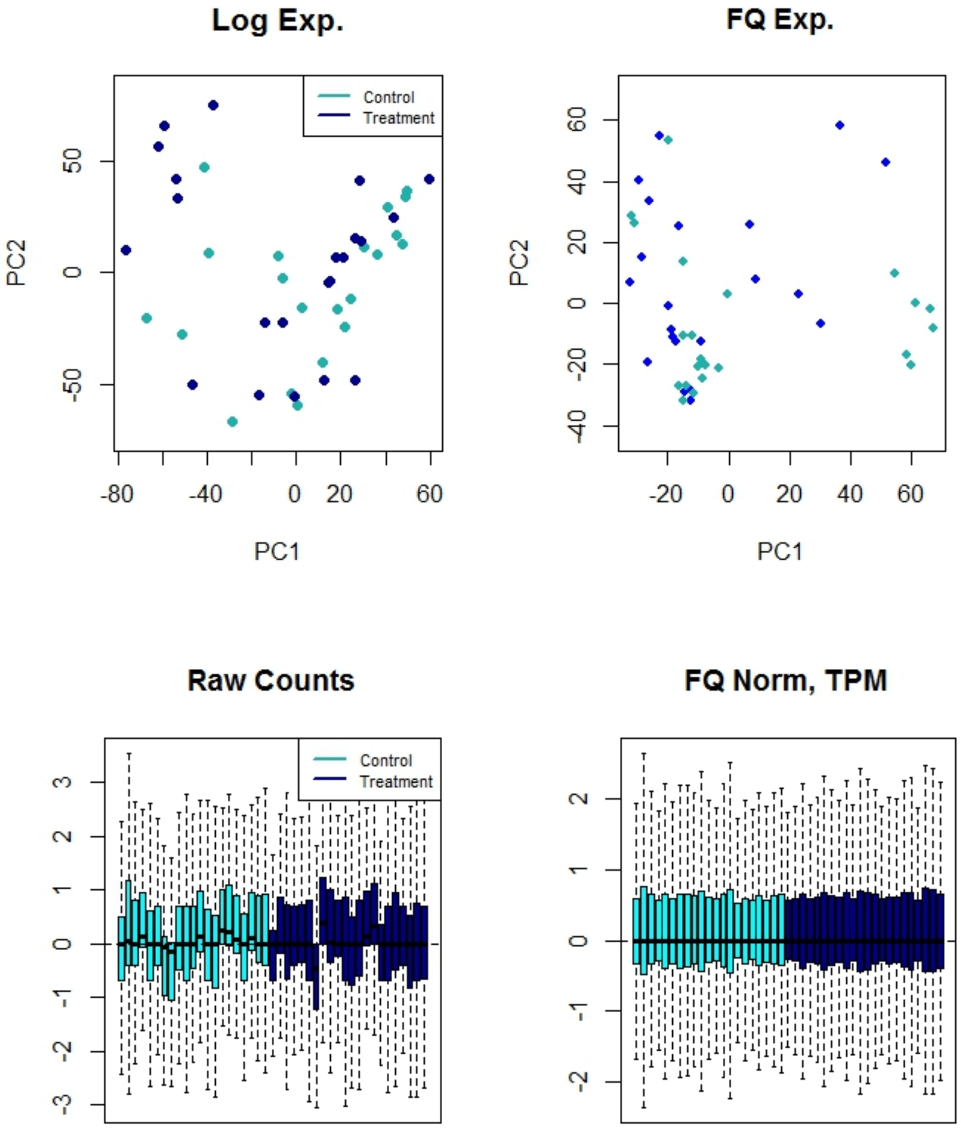

Spearman Correlation Clustering, RUVg counts, Batch1

SIDEseq Measure, RUVg Counts, Batch 1

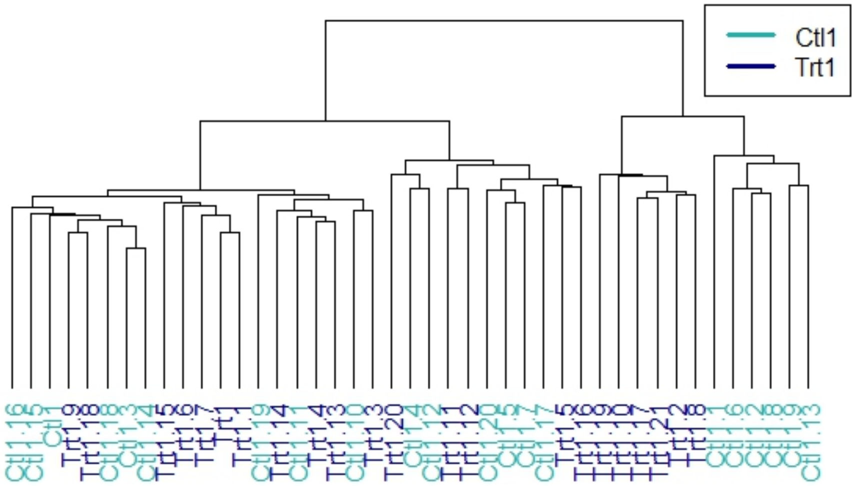
Supplementary Figure S5: After ‘RUVg’ normalization of the human ovarian cancer cells in batch one, hierarchical clustering using Spearman correlation and the SIDEseq measure fails to cluster cells by treatment and control status. It is hard to distinguish which method performs best on the ‘RUVg’ normalized counts for batch one.



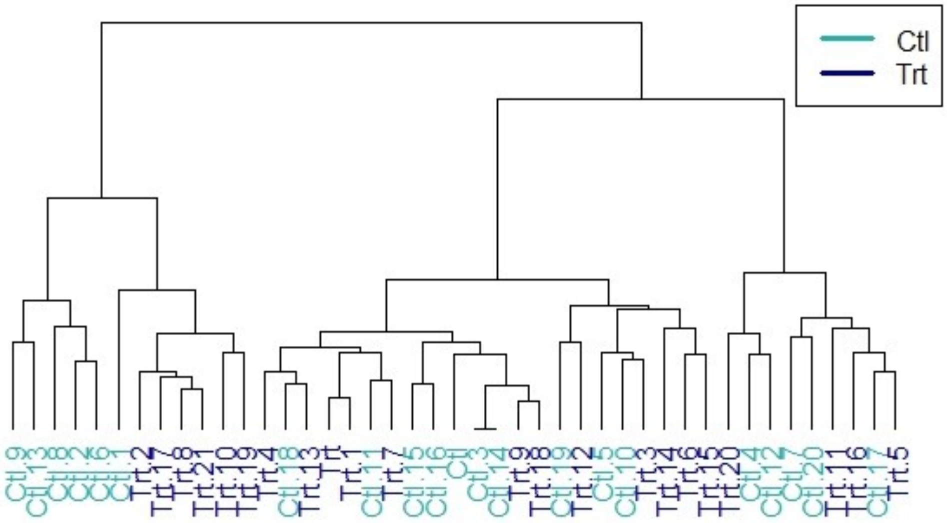
Supplementary Figure S6: Human ovarian cancer cells in batch two do not appear to cluster by treatment and control status after 'RUVg' normalization as well as the cells in batch one did. Un-normalized log expressions of cells in batch two were also used to plot the cells according to their first and second principal component for comparison (top left). This suggests that the thrombin treated cells did not differentiate from control cells as much as the TGFβ-1 treated cells in batch one. Cells in batch two do not cluster by treatment and control status when hierarchical clustering is used with various distance measures (figures not shown here).



Supplementary Figure S7: Log TPM expressions for the human ovarian cancer cells in batch two were full quantile normalized and then plotted according to their first and second principal component. Un-normalized log expressions of cells in batch one were also used to plot the cells according to their first and second principal component for comparison (top left). There is not an obvious improvement in the clustering of cells by treatment and control status after full quantile normalization.

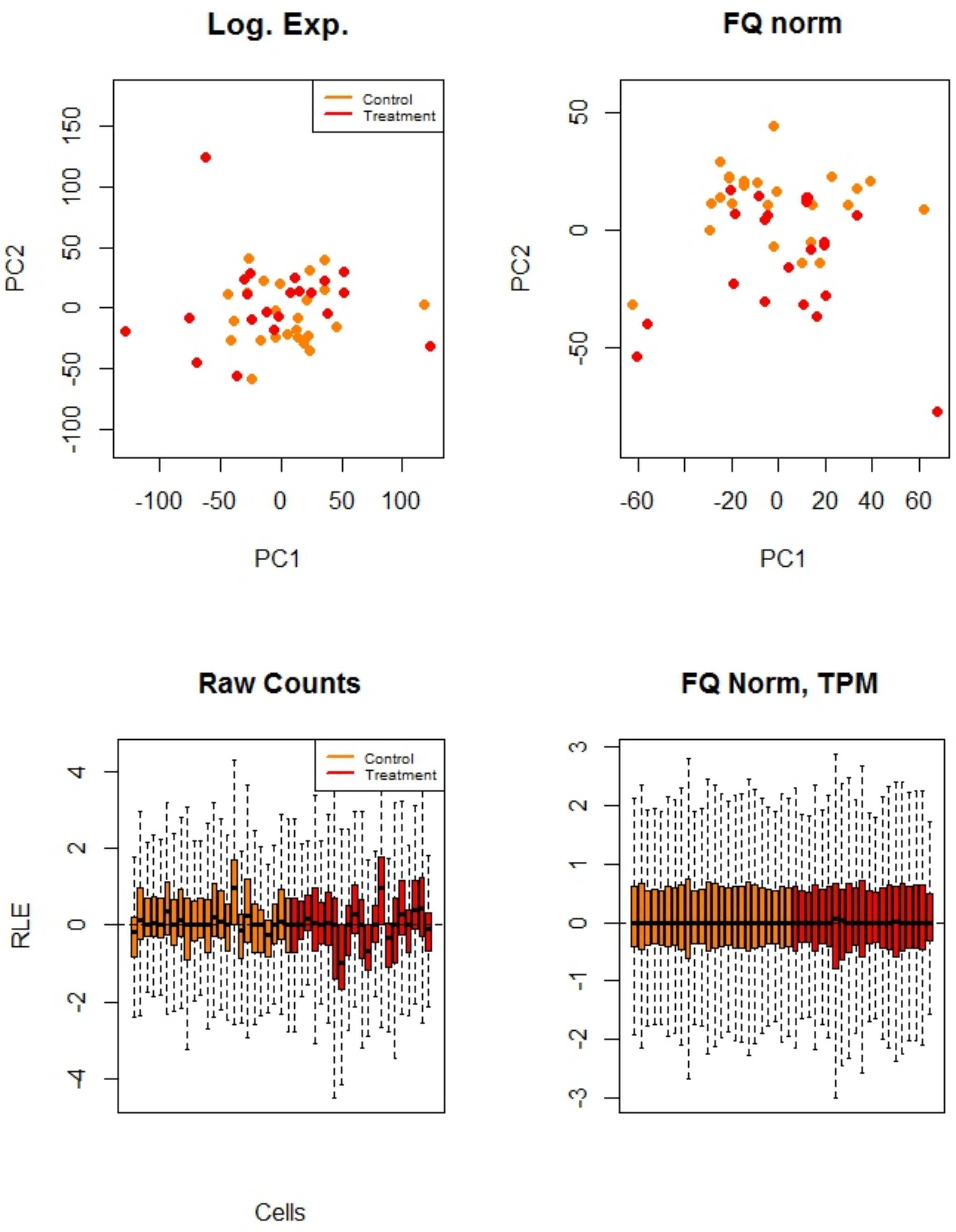


Spearman Correlation, FQ Normalization, Batch 1

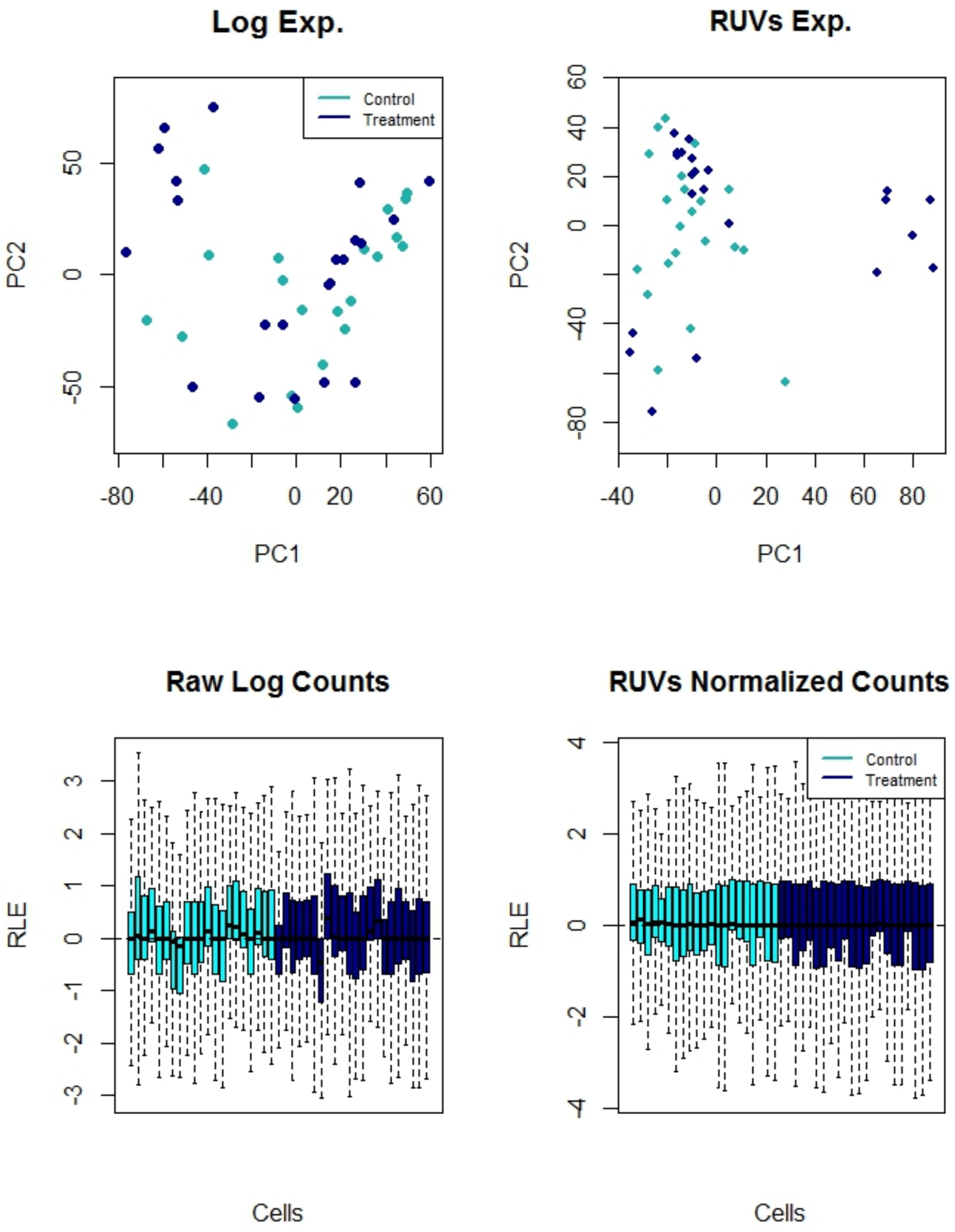


SIDEseq Measure, FQ Norm., Batch 1

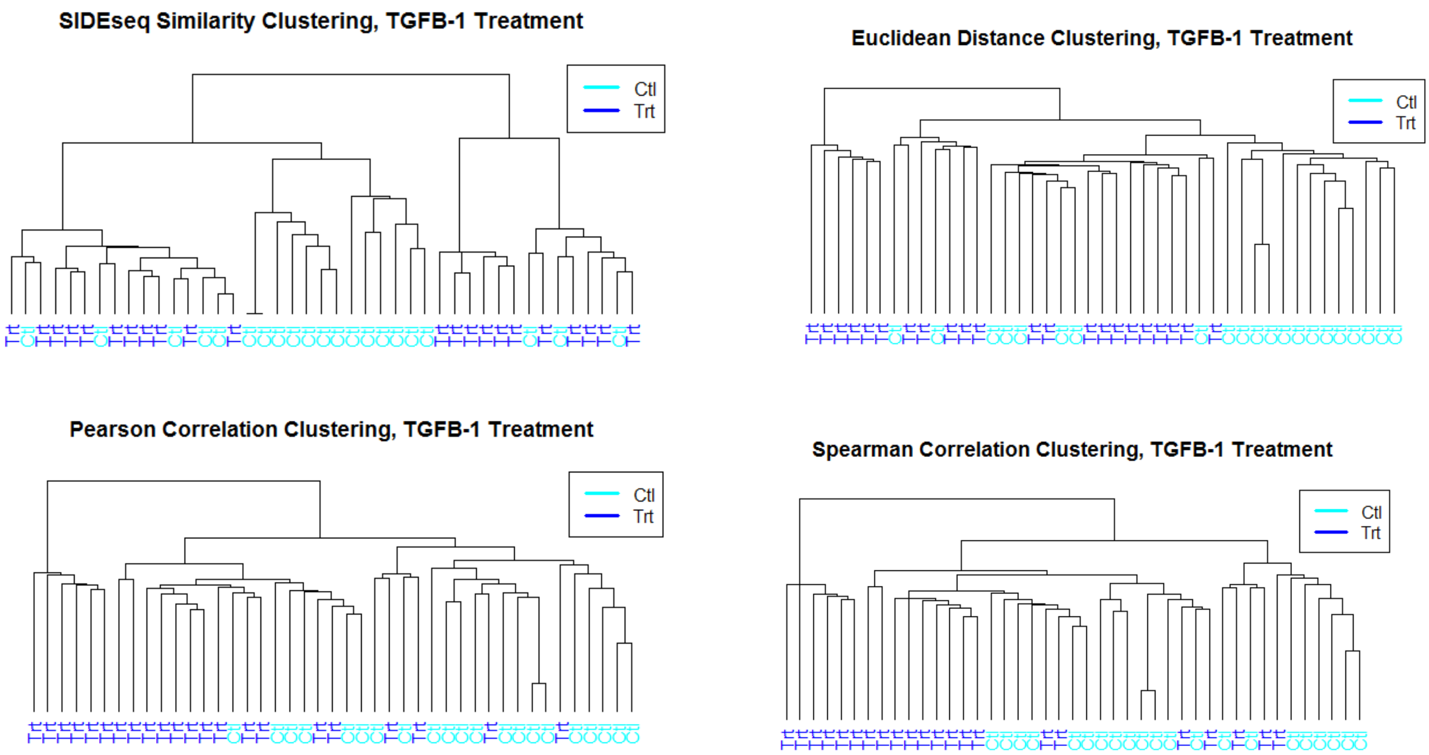
Supplementary Figure S8: After full quantile normalization of the human ovarian cancer cells in batch one, cells do not cluster by treatment and control status except for a small group of control cells and treatment cells. Clustering results are similar when Spearman correlation and the SIDEseq measure are used for hierarchical clustering.



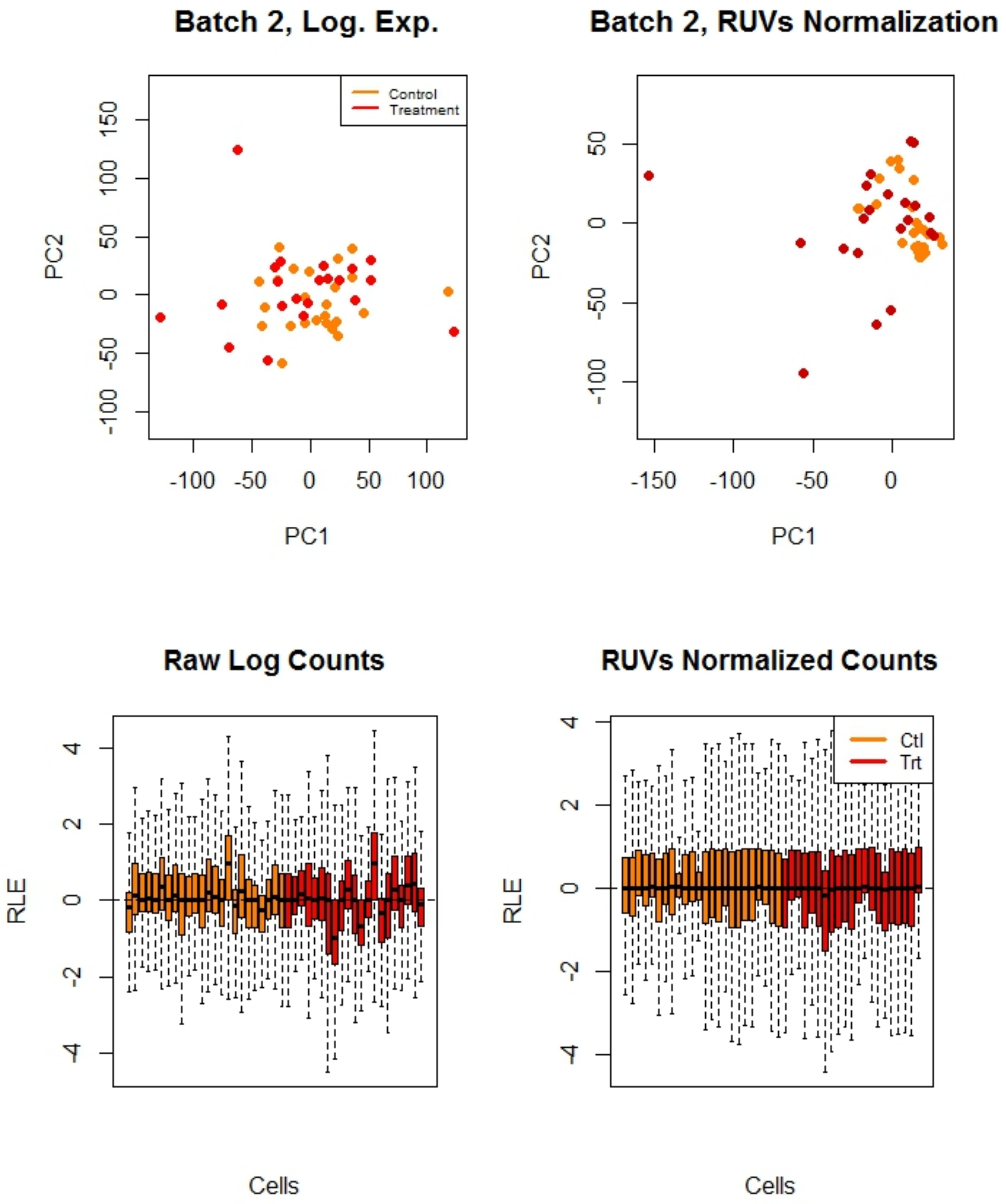
Supplementary Figure S9: When plotted according to their first two principal components, the human ovarian cancer cells in batch two do not separate by treatment and control status after full quantile normalization. Un-normalized log expressions of cells in batch two were also used to plot the cells according to their first and second principal component for comparison.



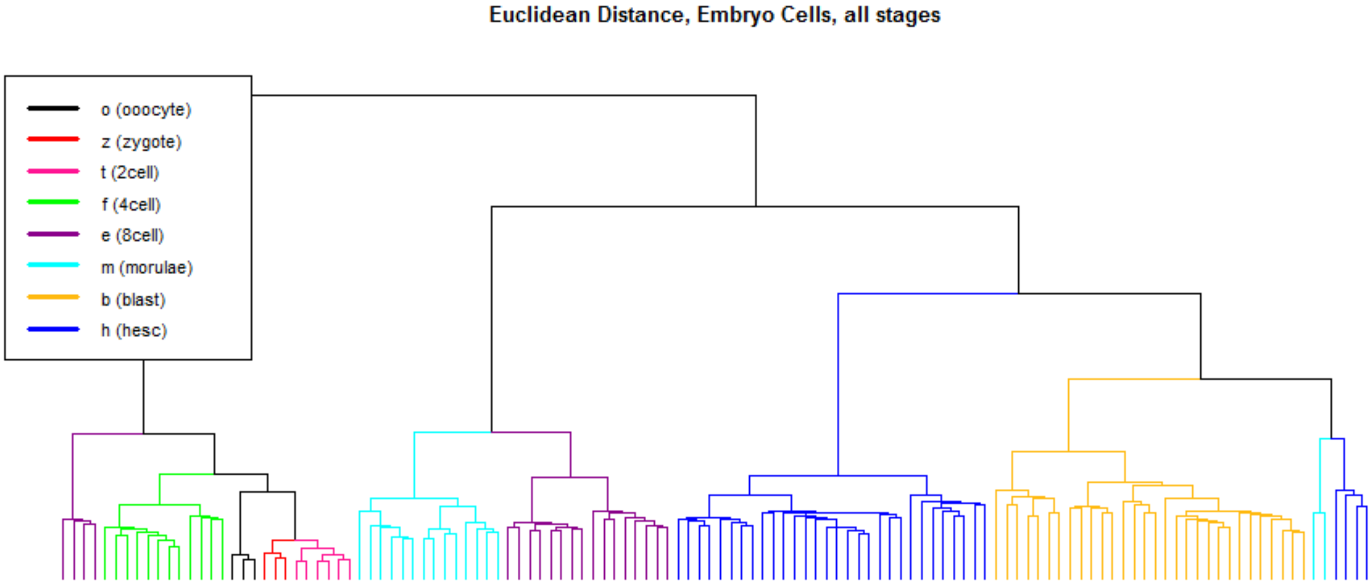
<u>Supplementary Figure S10:</u> After 'RUVs' normalization of the human ovarian cancer cells in batch one, there is an improvement in clustering of cells by treatment and control status. Un-normalized log expressions of cells in batch one were also used to plot the cells according to their first and second principal component for comparison.



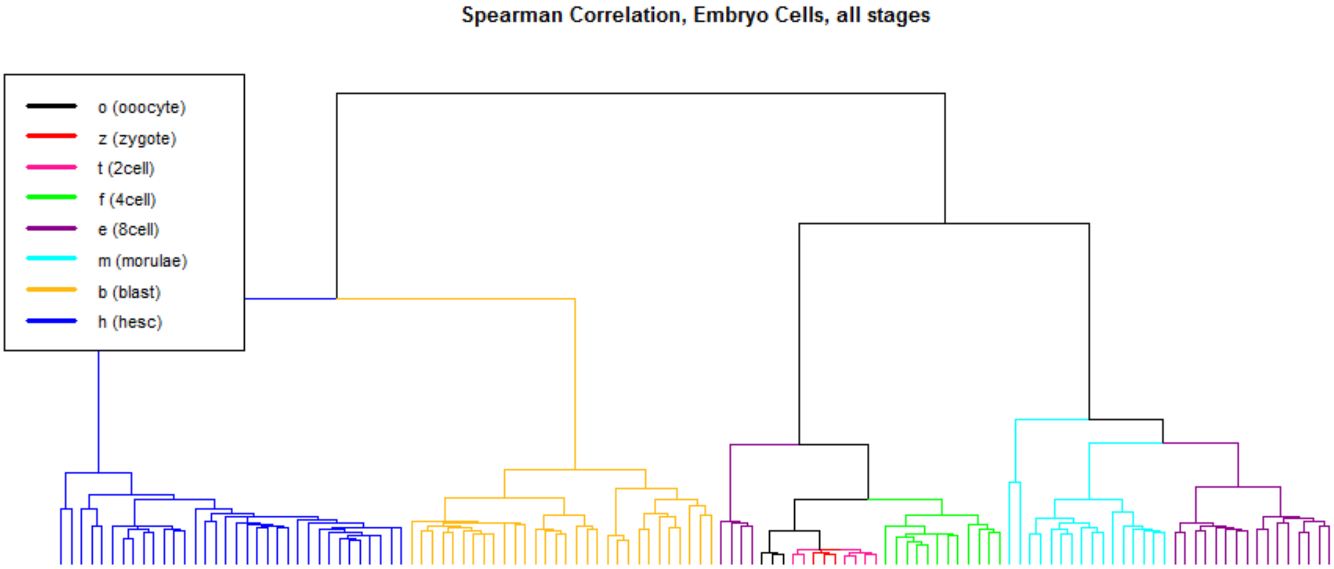
Supplementary Figure S11: After 'RUVs' normalization, the SIDEseq measure outperforms Spearman correlation when used with hierarchical clustering in terms of its ability to cluster treatment and control cells. Hierarchical clustering using the SIDEseq measure results in a large cluster of control cells and a large cluster of almost entirely treatment cells. There is a cluster which is a mixture of treatment and control cells which is positioned next to the cluster of all control cells. This suggests that this is a cluster with treatment cells which are in earlier an earlier stage of EMT or just began to transition, and so they cluster with some of the control cells. Pearson correlation results are very similar to Spearman correlation. Clustering with Euclidean distance results in a cluster of treated cells and a large cluster of control cells.



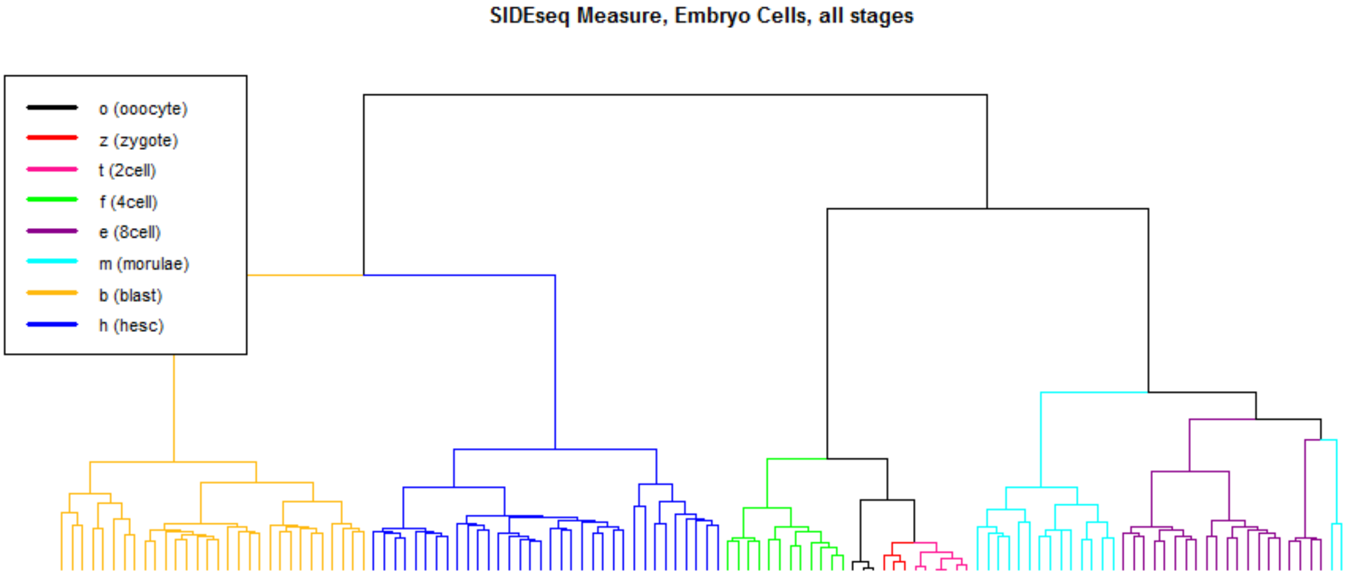
Supplementary Figure S12: When plotted according to their first and second principal components, the human ovarian cancer cells in batch two do not appear to cluster better by treatment and control status after ‘RUVs’ normalization. Un-normalized log expressions of cells in batch two were also used to plot the cells according to their first and second principal component for comparison. These figures suggest that the thrombin treatment cells in batch two were unable to differentiate from the control cells.



Supplementary Figure S13: When Euclidean distance is used for hierarchical clustering of the 124 human embryo cells from Yan *et al.* (2012), cells cluster relatively well by developmental stage. Although visually there are small places in the dendrograms where Euclidean distance is outperformed by Spearman correlation and SIDEseq (see Supplementary Figures S14 and S15), its ARI values are comparable to, and sometimes larger than those of Spearman correlation and SIDEseq (see Table 2 of the paper).



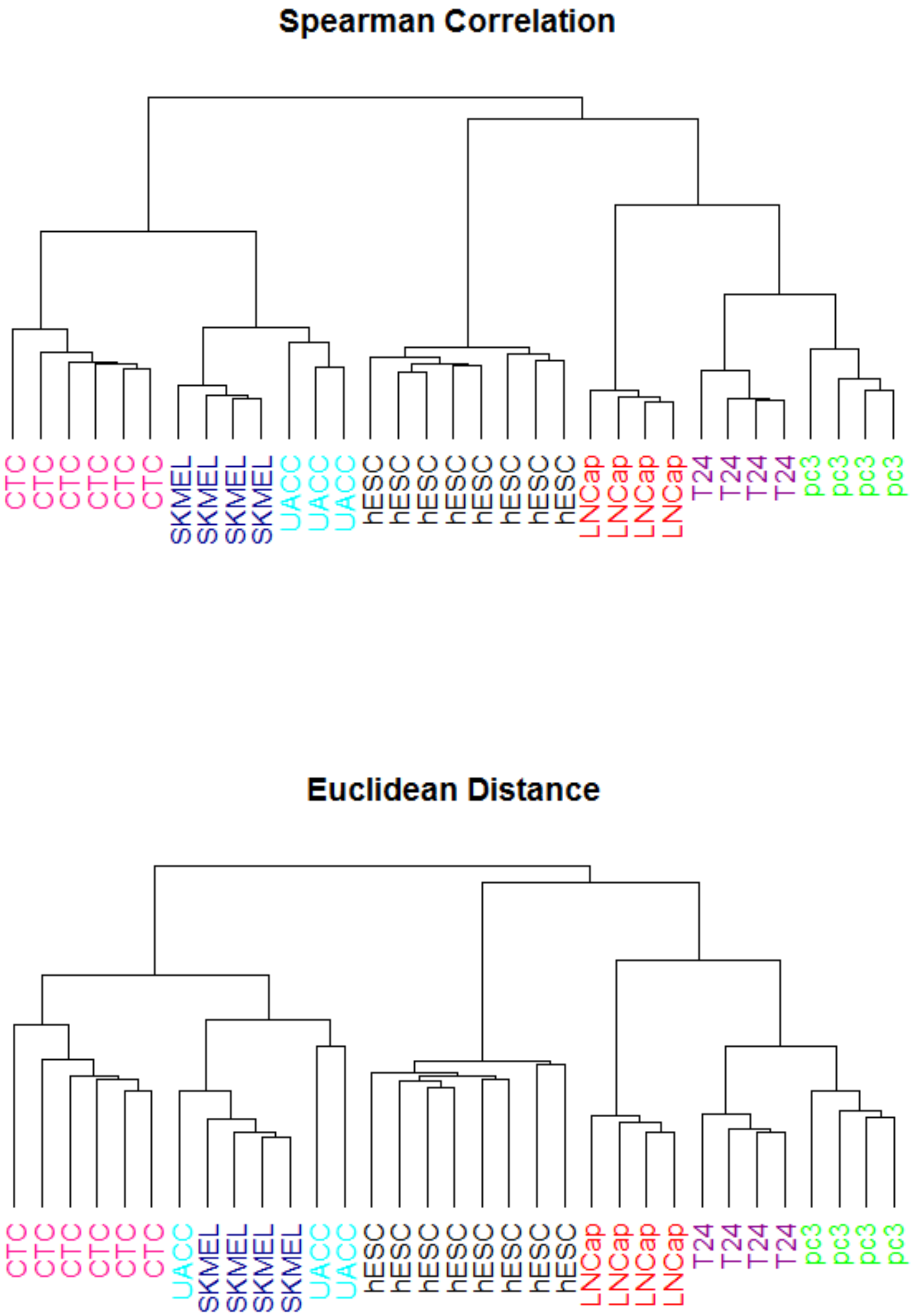
Supplementary Figure S14: The 124 human embryo cells from Yan *et al.* (2012) cluster well by developmental stage when Spearman correlation is used. However, two 2-cell stage cells cluster outside of a cluster of 2-cell and zygote cells. Furthermore, 2 morula cells cluster outside of a cluster of morula and 8-cell stage cells, and a group of four 8-cell stage cells cluster with the cells in the earlier stages of development. Results are similar when Pearson correlation is used. In terms of ARI values, Spearman correlation and the SIDEseq measure are very comparable, with SIDEseq having a slightly larger ARI value when the trees are cut to 9 clusters, due to the errors in the Spearman correlation clustering pointed out above.



Supplementary Figure S15: Hierarchical clustering using the SIDEseq measure of the 124 human embryo cells from Yan *et al.* (2012). Cells cluster perfectly by developmental stage when the proposed distance measure is used, except for two morula cells which cluster with the 8-cell stage cells. Note that the early developmental stages (oocyte to 2-cell) cluster perfectly, unlike with Spearman, Pearson correlation and Euclidean distance. In terms of ARI values, it is slightly outperformed by Euclidean distance when the dendrograms are cut to seven and eight clusters.

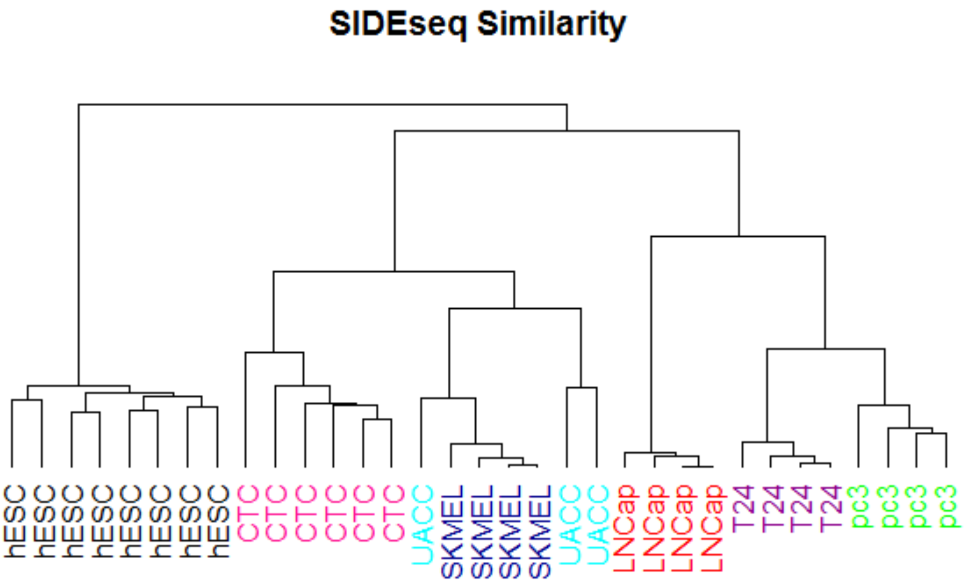
Clustering was also done on a data set from Ramsköld *et al*. (2012) who used a single-cell RNA-Seq platform called Smart-Seq. From the Gene Expression Omnibus (GEO) database, we downloaded RPKM gene expressions for human embryonic stem cells hESC (8 cells), prostate cancer cell lines LNCap (4 cells) and PC3 (4 cells), putative melanoma CTCs (6 cells) from peripheral blood, melanoma cell lines SKMEL5 (4 cells) and UACC257 (3 cells), and bladder cancer cell line T24 (4 cells). Following the filtering method of Xu *et al.* (2015) and Ramsköld *et al*. (2012), we discarded genes with average RPKM less than 20 across all thirty-three cells, leaving over three thousand genes. Xu *et al*. tested the same clustering methods on this data set as they did for the human embryo cells. When Pearson and Spearman correlation were used for hierarchical clustering, the clustering results improved upon all methods tested by Xu *et al*. except their SNN-Cliq algorithm, which was matched in performance by the hierarchical clustering. Euclidean distance and the SIDEseq similarity gave very similar results when used to cluster these cancer cells, performing only slightly worse than Pearson and Spearman correlation.

		Adjusted Rand Index			
Public Data Set	**# Clusters**	**Pearson Correlation**	**Spearman Correlation**	**Euclidean Distance**	**SIDEseq Similarity**
Cancer Cells, Ramsköld *et al*.	7	1	1	0.794	0.951

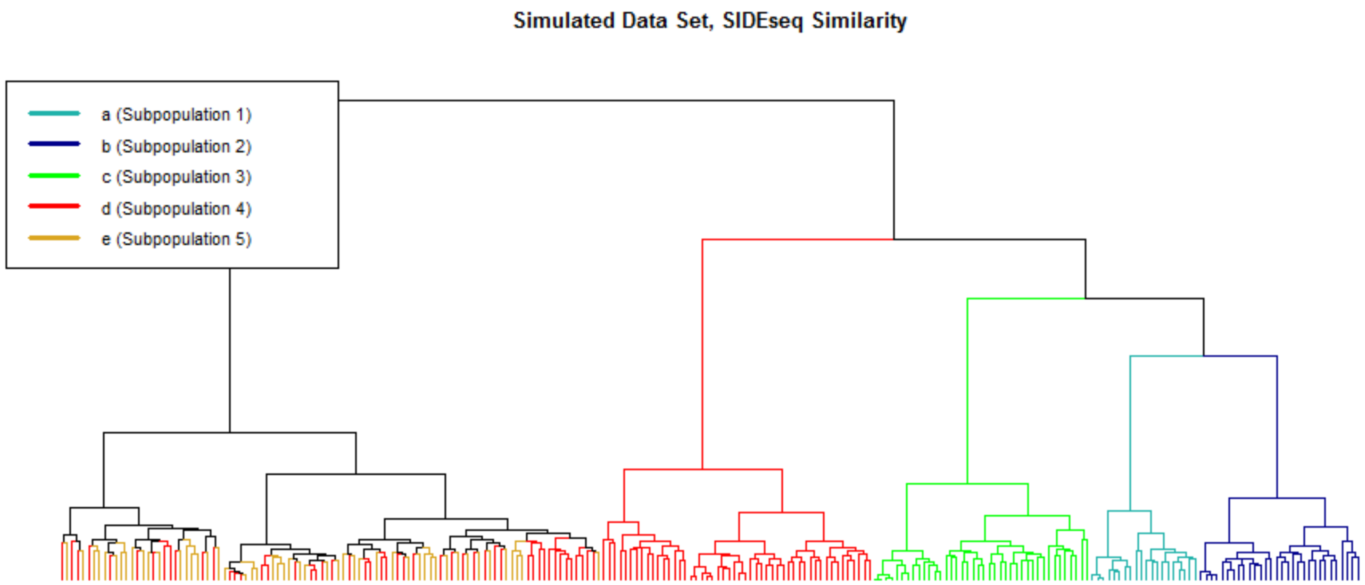

Spearman Correlation

Euclidean Distance

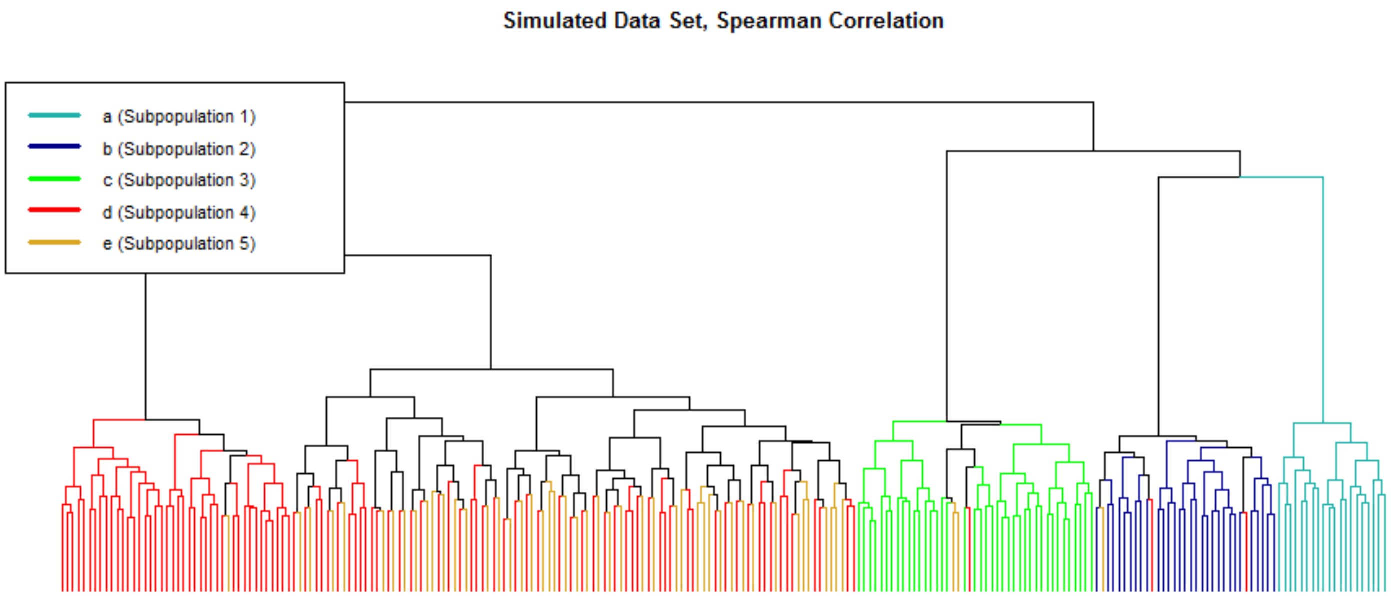
Supplementary Figure S16: Euclidean distance and Spearman correlation have similar results when used to do hierarchical clustering of 33 human cancer cells and embryo cells from Ramsköld *et al*. (2012).

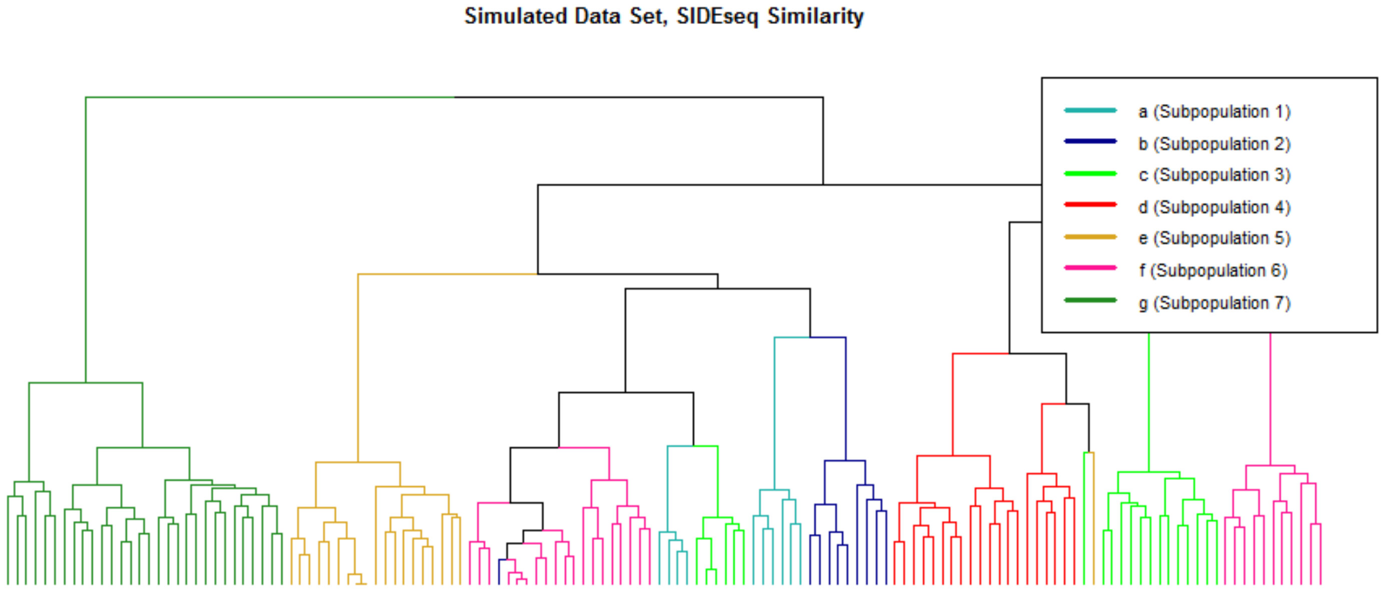


Supplementary Figure S17: When the SIDEseq measure is used for hierarchical clustering of 33 human cancer cells and embryo cells from Ramsköld *et al*. (2012), results are similar to the Euclidean distance and Spearman correlation results

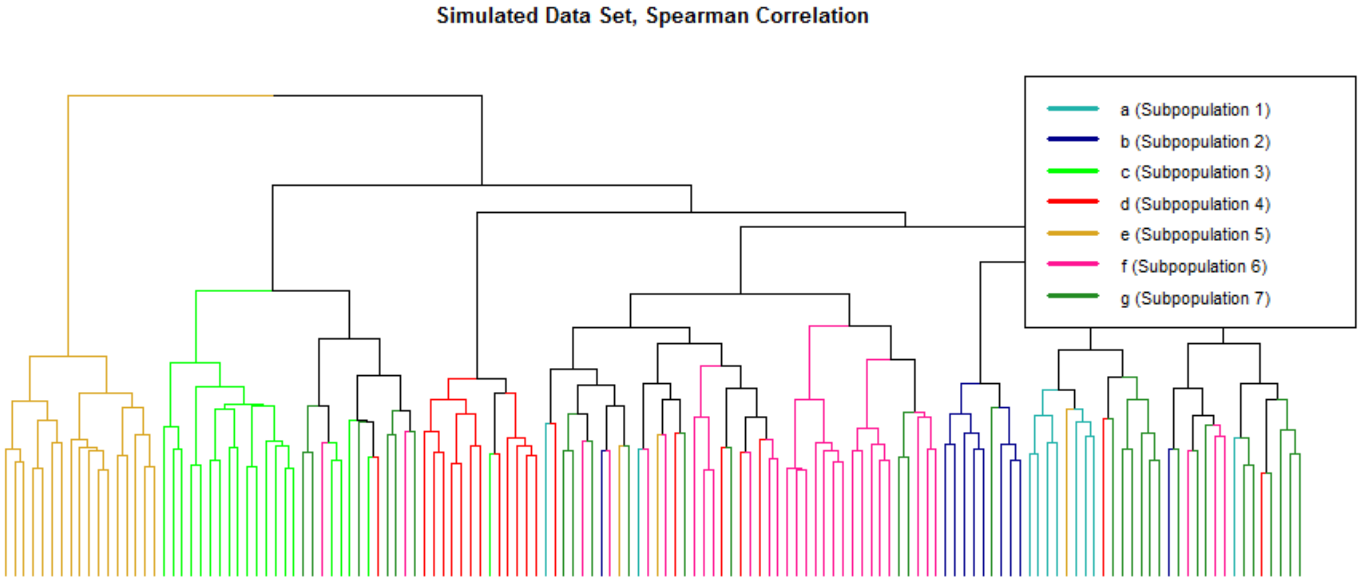


<u>Supplementary Figure S18:</u> In this simulated data set, five subpopulations of cells all have different probabilities that a gene will be differentially expressed. The SIDEseq measure results in an ARI of 0.552, and Spearman correlation results in an ARI of 0.500.





<u>Supplementary Figure S19:</u> For this simulated data set, 7 subpopulations of cells have a lower degree of differential expression. The SIDEseq similarity measure has an ARI of 0.729, while Spearman correlation has an ARI of 0.405.



References for supplementary material:

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Ramsköld, D., et al. (2012) Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells., *Nature biotechnology*, 30, 777-782.

Risso, Davide et al. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples., *Nature Biotechnology*, 32.9, pp. 896–902.

Yan, L. et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells., *Nature Structural and Molecular Biology,* 20, pp. 1131–1139.