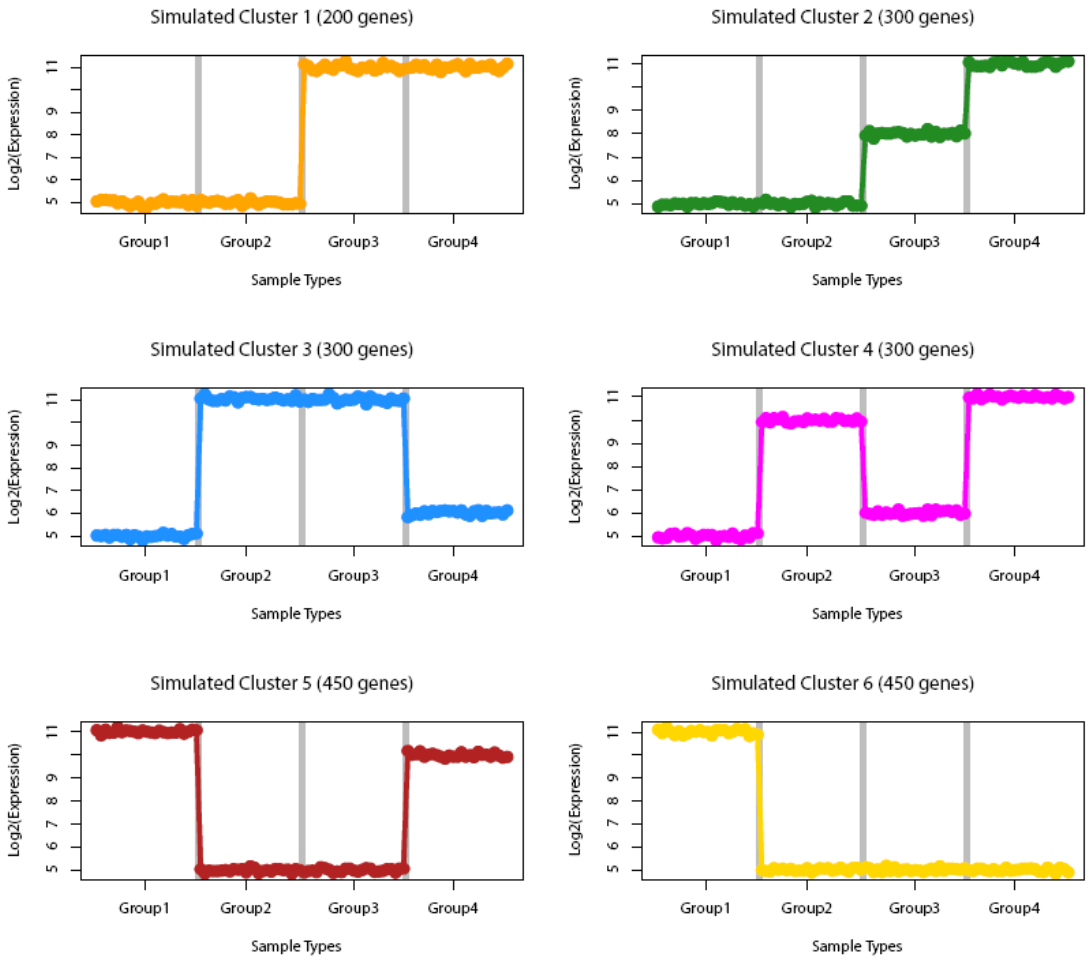
Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data for a Large Number of Genes with a Moderate Number of Samples

We simulated gene expression data for 2000 genes and 100 samples, corresponding to 25 replicates within each of four phenotypic groups. Using different mean values, we simulated six clusters in this data set under a Normal distribution, using a fixed standard deviation value of 1.5 throughout.



Using complete linkage agglomerative hierarchical clustering with Pearson correlation, the maximum number of clusters to split the genes into clusters that had at least five genes, was 10. We used the following metrics to test the most appropriate number of clusters between the interval of 2 to 10. The informativeness metric was the only metric able to recover the correct number of clusters in which this data set was simulated under, as illustrated in the table below.

	Compactness Metrics				Stability Metrics					
Number of Simulated Clusters	**Gap Statistic**	**Connectivity**	**Dunn Index**	**Silhouette Width**	**APN**	**AD**	**ADM**	**FOM**	**Modified F statistic**	**Informativeness Metric**
6	5	2	5	4	2	10	2	10	2	6

Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data Using Hierarchical Clustering

We simulated gene expression data sets for 300 genes and 100 samples, corresponding to 25 replicates within each of four phenotypic groups. Using different combinations of mean values, data were simulated from a Normal distribution, with a fixed standard deviation value of 1.5 throughout. 100 data sets were generated for four, six and eight clusters and were used to test the performance of the ten metrics. All ten metrics were evaluated over an interval whose upper limit was defined by the maximum number of clusters that gave clusters with at least five genes.

We used both complete-linkage agglomerative hierarchical clustering and k-means clustering algorithms, however as shown later in this section, the performance of k-means was so extremely poor that the original simulated data sets had to be revised for our testing purposes.

Here, we first report on the results obtained with hierarchical clustering.

In all instances, the informativeness metric on average was able to estimate the number of clusters correctly and performed similarly to the Gap statistic.

Table 1: Performance of 10 Metrics on the 4-Cluster Data Sets

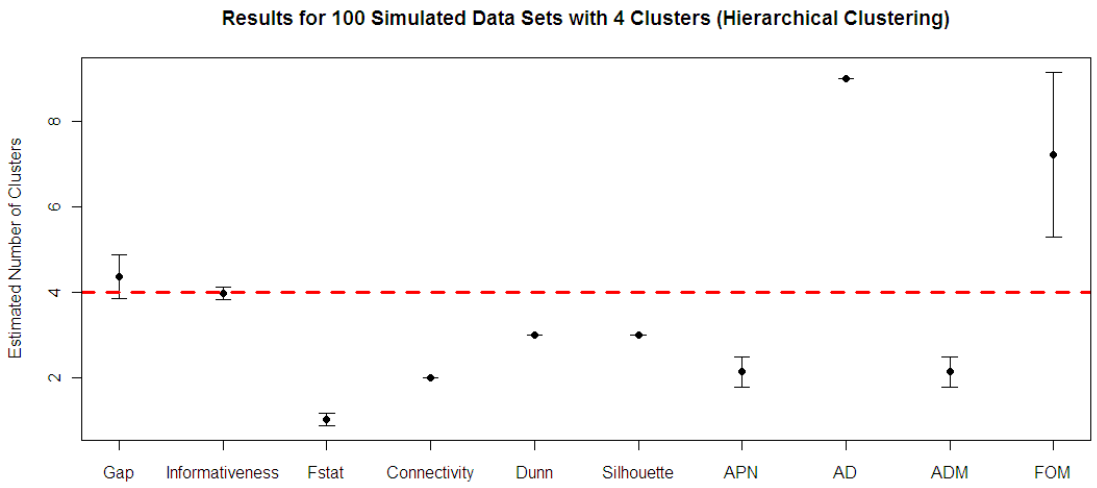
K=4	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	4.37	3.98	1.02	2.00	3.00	3.00	2.14	9.00	2.14	7.22
Standard Deviation	0.506	0.141	0.141	0	0	0	0.349	0	0.349	1.94
Range	[3, 5]	[3, 4]	[1, 2]	[2, 2]	[3, 3]	[3, 3]	[2, 3]	[9, 9]	[2, 3]	[4, 9]

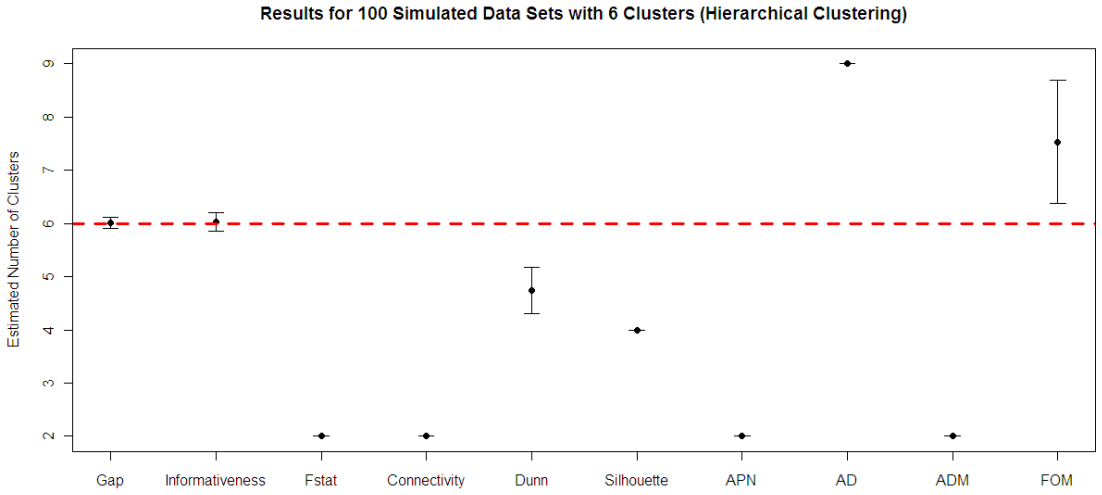
Table 2: Performance of 10 Metrics on the 6-Cluster Data Sets

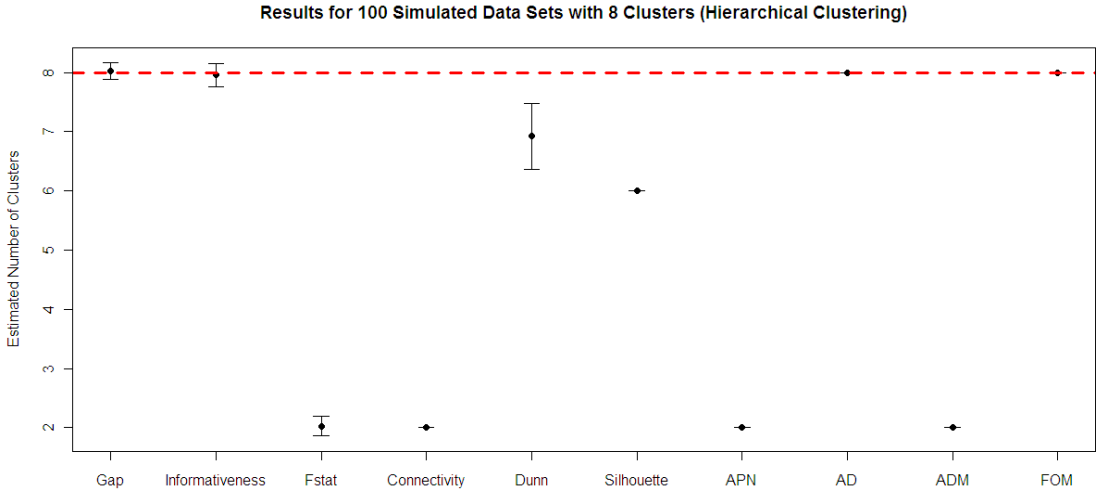
K=6	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	6.01	6.03	2.00	2.00	4.74	4.00	2.00	9.00	2.00	7.53
Standard Deviation	0.100	0.171	0	0.441	0	0	0	0	0	1.16
Range	[6, 7]	[6, 7]	[2, 2]	[2, 2]	[4, 5]	[4, 4]	[2, 2]	[9, 9]	[2, 2]	[6, 9]

Table 3: Performance of 10 Metrics on the 8-Cluster Data Sets

K=8	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	8.02	7.96	2.03	2.00	6.92	6.00	2.00	8.00	2.00	8.00
Standard Deviation	0.141	0.197	0.171	0	0.563	0	0	0	0	0
Range	[8, 9]	[7, 8]	[2, 3]	[2, 2]	[3, 7]	[6, 6]	[2, 2]	[8, 8]	[2, 2]	[8, 8]

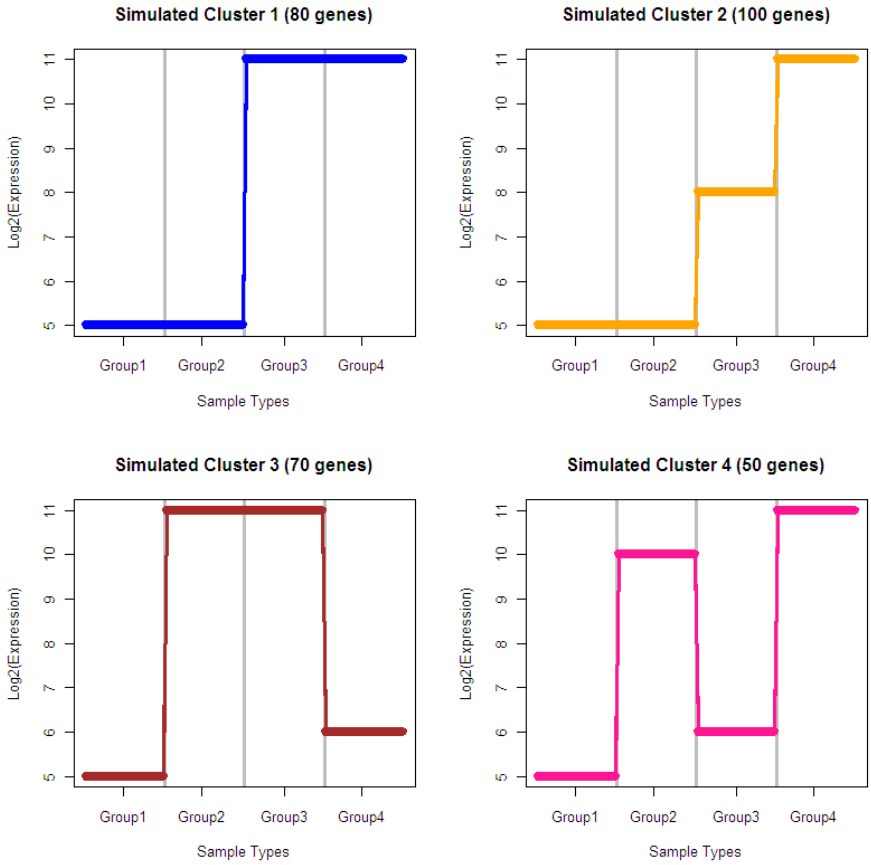

Results for 100 Simulated Data Sets with 4 Clusters (Hierarchical Clustering)
Estimated Number of Clusters
2
4
6
8
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM


Results for 100 Simulated Data Sets with 6 Clusters (Hierarchical Clustering)
Estimated Number of Clusters
2
3
4
5
6
7
8
9
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM


Results for 100 Simulated Data Sets with 8 Clusters (Hierarchical Clustering)
Estimated Number of Clusters
2
3
4
5
6
7
8
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM

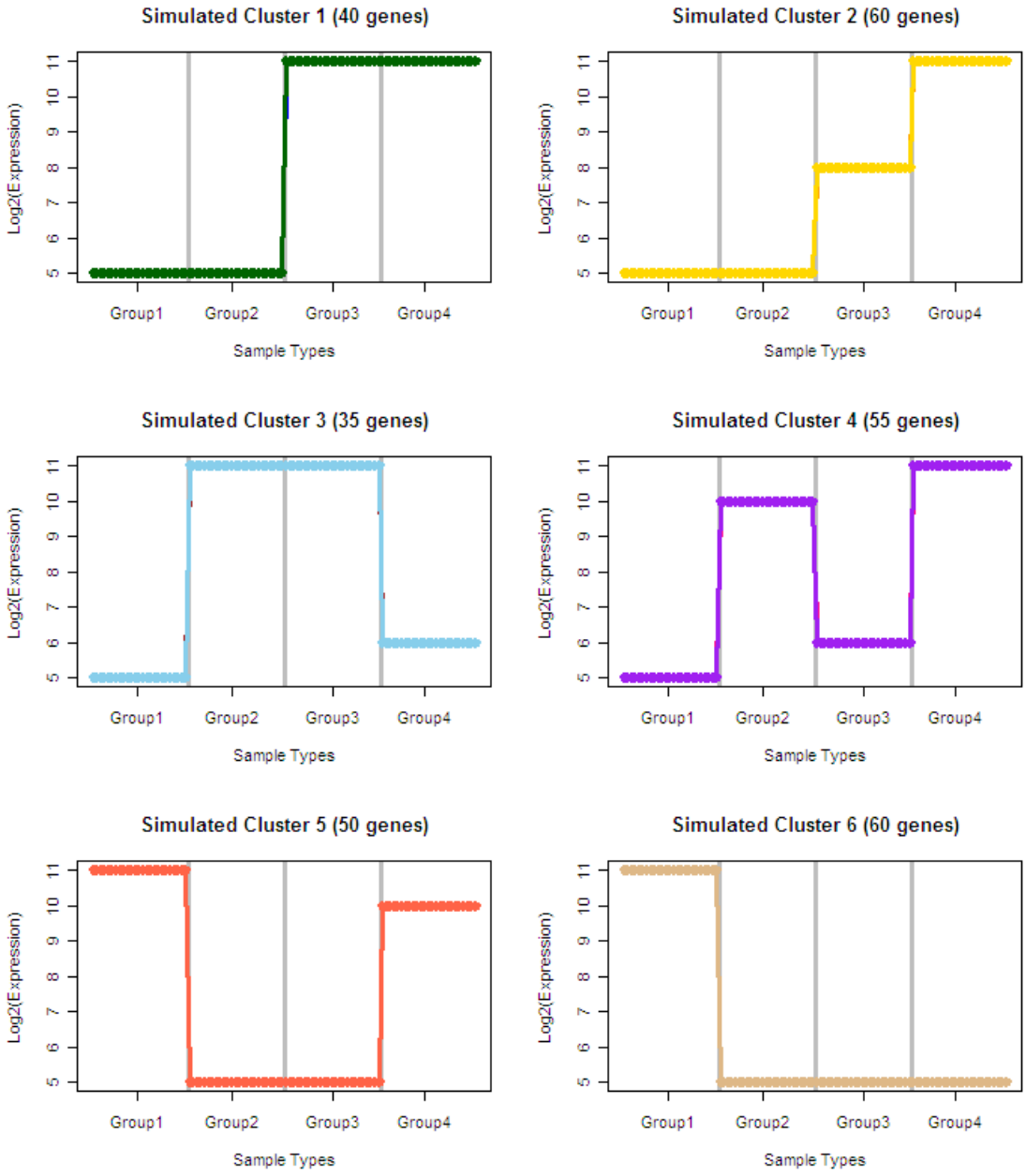
4-Cluster Data Set Reference Clusters

100 data sets were simulated using the following set of mean values to make up the four artificial clusters:



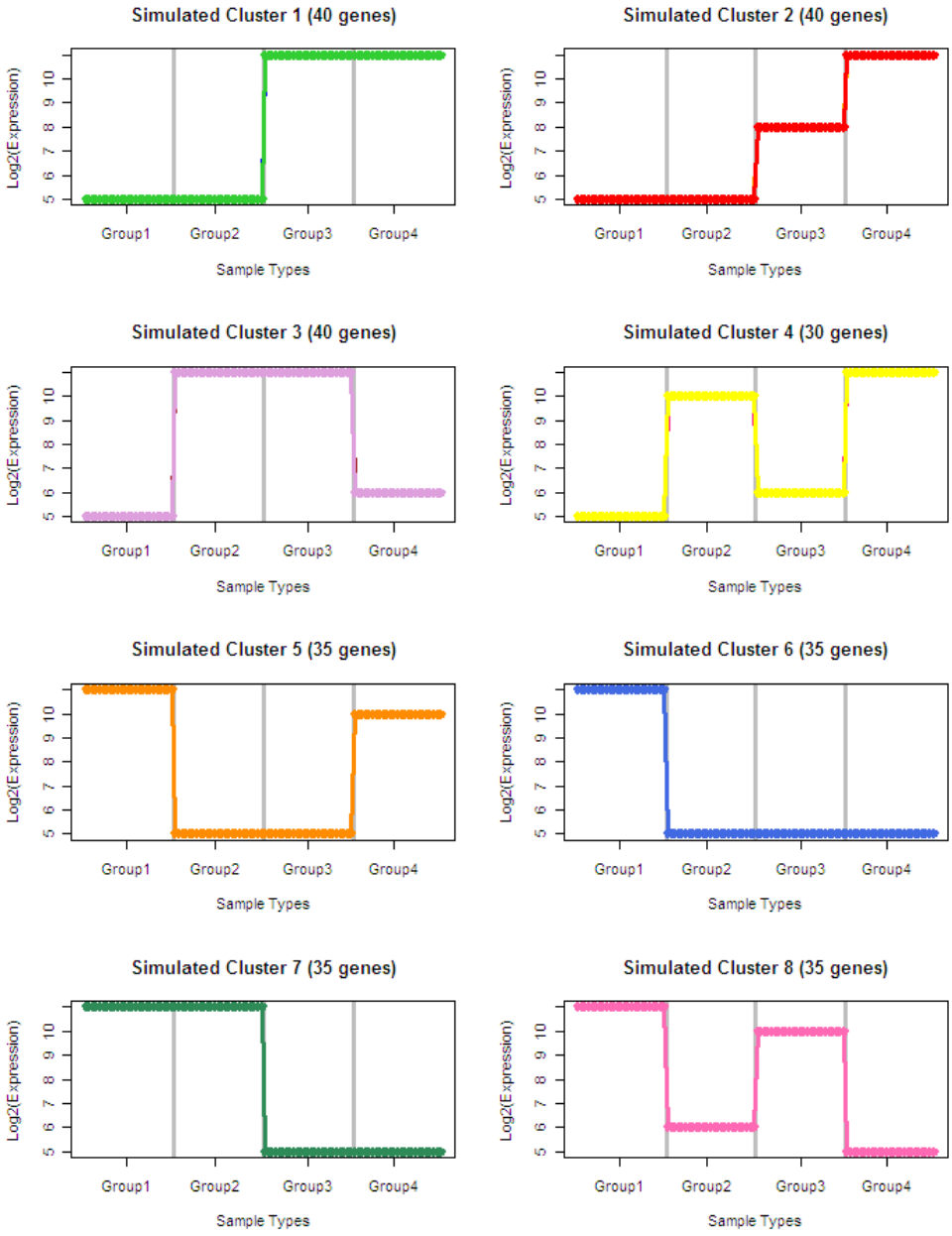
6-Cluster Data Set Reference Clusters

100 data sets were simulated using the following set of mean values to make up the six artificial clusters:



8-Cluster Data Set Reference Clusters

The 100 data sets were simulated using the following set of mean values to make up the eight artificial clusters:



Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data Using k-Means Clustering

Using the same 100 simulated data sets for known numbers of clusters (four, six and eight) that had been tested using hierarchical clustering, we applied a k-means clustering algorithm using Pearson's correlation coefficient as the similarity metric. We used the Kmeans implementation in the amap package, with parameters max.iters = 300 and nstart = 300. The test interval was specified as before, where the upper limit was defined by the maximum number of clusters with all clusters containing at least five genes.

With this approach, it was evident that for the majority of simulations for data sets with four, six and eight clusters, the quality of the clusters produced by the k-means clustering algorithm was extremely poor. One way this is apparent is by assessing the interval that was tested; the majority of test intervals for the 100 data sets did not include the true number of clusters the data was simulated under. Therefore, data sets with test intervals that did not contain this true value were unable to be used for testing the performance of the ten metrics since it is impossible under this setting for the metric to recover the true number of clusters.

	K = 4	K = 6	K = 8
Number of Analyses without the True Value of K	82	83	66
Number of Simulated Data Sets that Can Be Used for Testing	**18**	**17**	**34**

Since the poor quality observed is a direct result of the clustering algorithm (k-means) and not related to the metrics selected – we proceeded with our testing of these metrics in two ways: first, we ran tests on the simulations whose test interval did contain the true number of clusters since these results are useful as a consistent comparison to the results obtained from hierarchical clustering; second, we altered the reference clusters to a much simplified version and modified the way in which estimates for the number of clusters was tested. Results from the former are reported in Part 1, the latter appear in Part 2 of this supplementary section.

Part 1: Using A Standard Set of Simulated Data Sets

Overall the FOM, AD, informativeness metric and the Gap statistic do an excellent job at estimating the correct number of clusters on average. We must however interpret these results in the context that they are based off performance in 18 simulated data sets (for the 4-cluster data set), 17 simulated data sets (for the 6-cluster data set) and 34 simulated data sets (for the 8-cluster data set).

Table 1: Performance of 10 Metrics on the 4-Cluster Data Sets

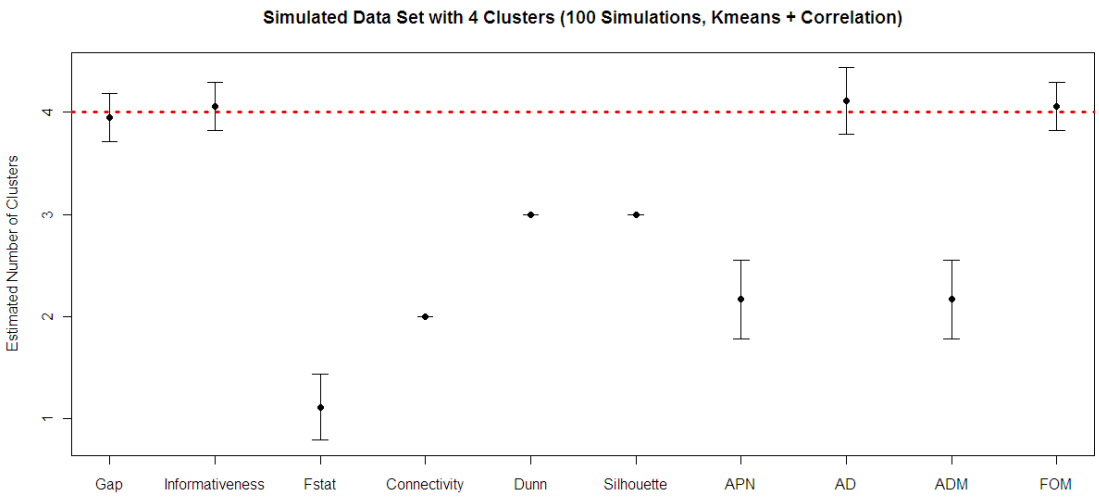
K=4	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	3.94	4.06	1.11	2.00	3.00	3.00	2.17	4.11	2.17	4.06
Standard Deviation	0.236	0.235	0.323	0	0	0	0.383	0.323	0.383	0.236
Range	[3. 4]	[4, 5]	[1, 2]	[2, 2]	[3, 3]	[3, 3]	[2, 3]	[4, 5]	[2, 3]	[4, 5]

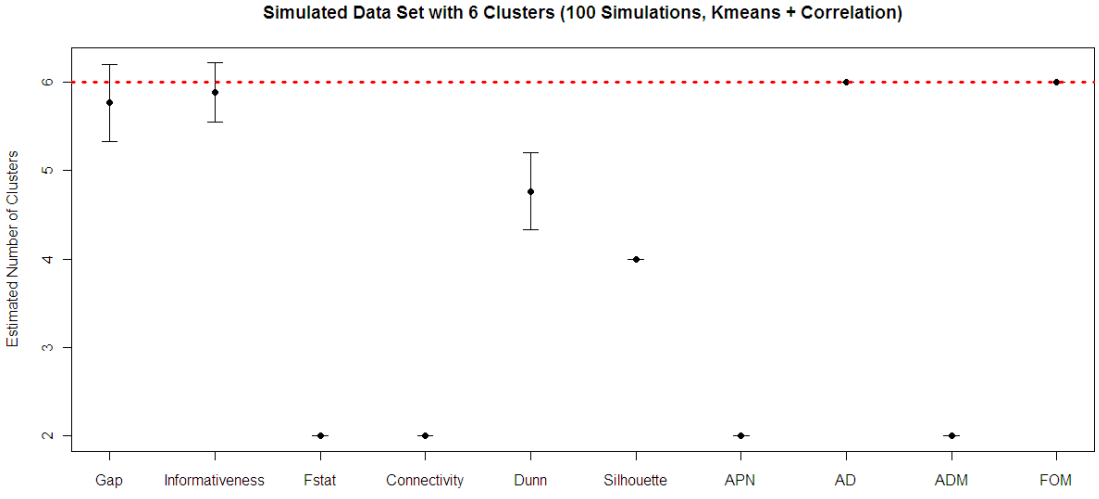
Table 2: Performance of 10 Metrics on the 6-Cluster Data Sets

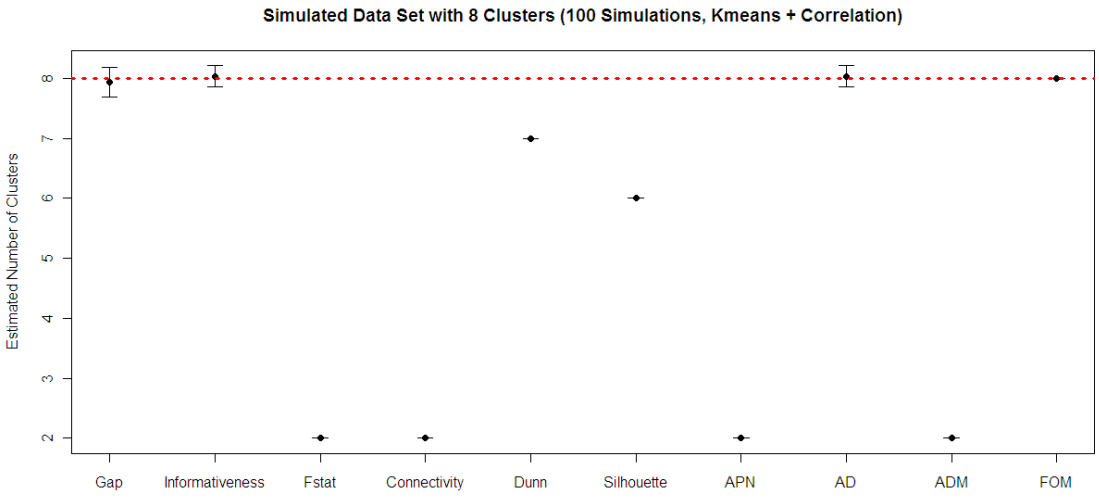
K=6	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	5.76	5.88	2.00	2.00	4.76	4.00	2.00	6.00	2.00	6.00
Standard Deviation	0.437	0.332	0	0	0.437	0	0	0	0	0
Range	[5, 6]	[5, 6]	[2, 2]	[2, 2]	[4, 5]	[4, 4]	[2, 2]	[6, 6]	[2, 2]	[6, 6]

Table 3: Performance of 10 Metrics on the 8-Cluster Data Sets

K=8	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	7.94	8.03	2.00	2.00	7.00	6.00	2.00	8.031	2.00	8.00
Standard Deviation	0.246	0.177	0	0	0	0	0	0.177	0	0
Range	[7, 8]	[8, 9]	[2, 2]	[2, 2]	[7, 7]	[6, 6]	[2, 2]	[8, 9]	[2, 2]	[8, 8]


Simulated Data Set with 4 Clusters (100 Simulations, Kmeans + Correlation)
Estimated Number of Clusters
1
2
3
4
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM


Simulated Data Set with 6 Clusters (100 Simulations, Kmeans + Correlation)
Estimated Number of Clusters
2
3
4
5
6
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM


Simulated Data Set with 8 Clusters (100 Simulations, Kmeans + Correlation)
Estimated Number of Clusters
2
3
4
5
6
7
8
Gap
Informativeness
Fstat
Connectivity
Dunn
Silhouette
APN
AD
ADM
FOM

Part 2: Using A Standard Set of Simulated Data Sets

To circumvent the problem of having so many fewer simulated data sets available for testing, we simulated data from a set of parameters that had artificial clusters with the simplest separation across the four phenotypic groups. Given their simplicity, these clusters are not accurate representations of the profiles we tend to see with real gene expression data however they provide us with a means to evaluate performance of our metrics with clusters that have been produced by an alternative clustering algorithm to hierarchical clustering.

We also altered the way in which the test interval was computed; instead of using an incremental search with a stopping criterion, we used an arbitrary fixed upper limit (ten for the 4-cluster and 6-cluster data sets, twelve for the 8-cluster data sets) and tested all candidates within the interval from two to the upper limit. As a result, for the 4-cluster and 6-cluster data sets, more of the simulations were useful for testing.

(Note: given the large number of possibilities for generating the 8-cluster data set, we did not invest more time in finding a set of artificial clusters whose patterns would result in a larger number of testable data sets with application of the k-means clustering).

	K = 4	K = 6	K = 8
Number of Analyses without the Simulated K	0	10	65
Number of Simulated Data Sets that Can Be Used for Testing	**100**	**90**	**35**

On average, all metrics struggled to estimate the true numbers of clusters. For the 4-cluster data sets, the Dunn index and Silhouette width metrics were the only ones to correctly estimate the number of clusters. The Gap statistic and the informativeness metric had the best average performance for the 6-cluster and 8-cluster data sets respectively.

ADM, APN, the connectivity score and the F-statistic consistently under-estimated the number of clusters for all 4-cluster, 6-cluster and 8-cluster data sets. For the 4-cluster and 6-cluster data sets, the Gap statistic, AD, FOM and informativeness metric produced over-estimates.

Because of the failure of k-means to generate testable clusters on the data sets that were originally simulated and coupled with the simplicity of the clusters represented here, the main conclusions we are able to draw from this analysis is that the informativeness metric demonstrated performance similar to the Gap statistic.

Table 1: Performance of 10 Metrics on the 4-Cluster Data Sets

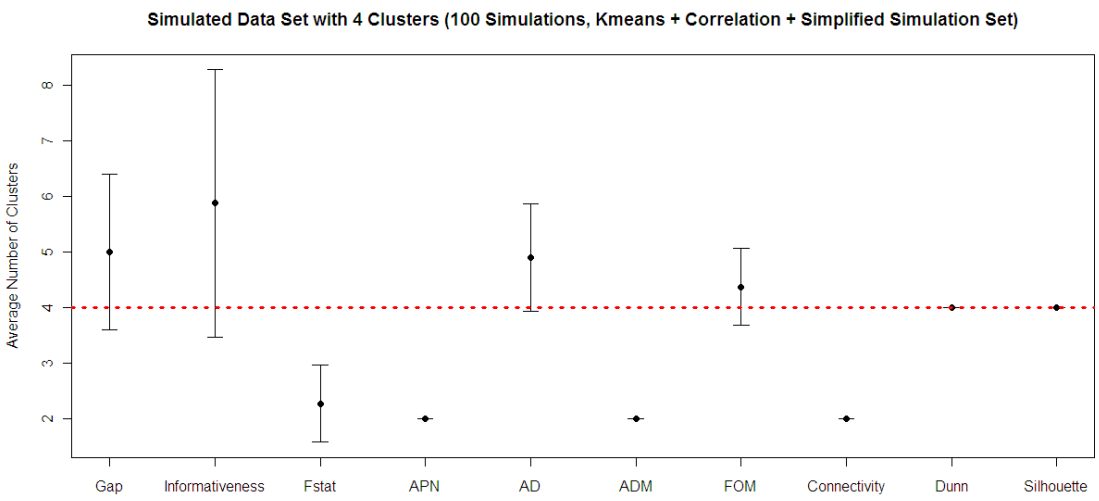
K=4	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	5.01	5.88	2.28	2.00	4.00	4.00	2.00	4.90	2.00	4.38
Standard Deviation	1.40	2.40	0.697	0	0	0	0	0.969	0	0.693
Range	[4,8]	[4,10]	[2,4]	[2,2]	[4,4]	[4,4]	[2,2]	[4,7]	[2,2]	[4,7]

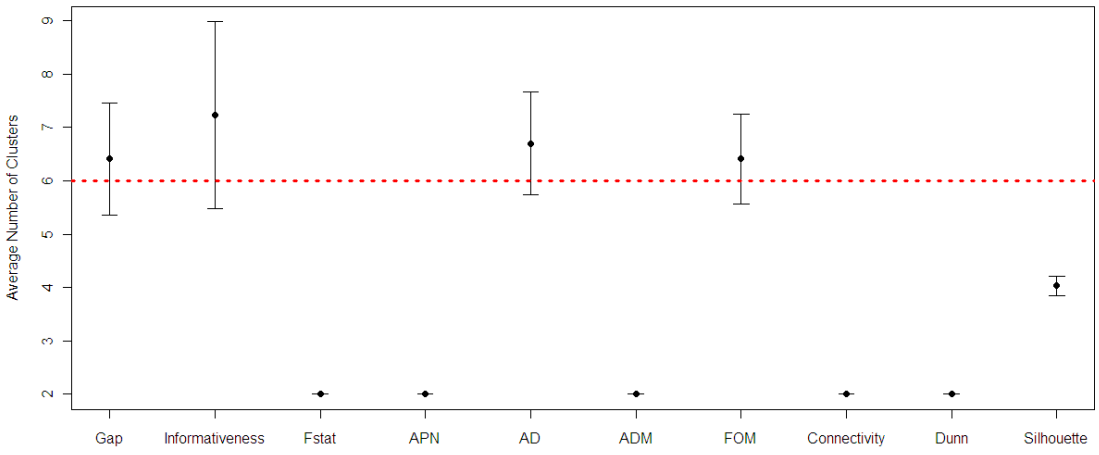
Table 2: Performance of 10 Metrics on the 6-Cluster Data Sets

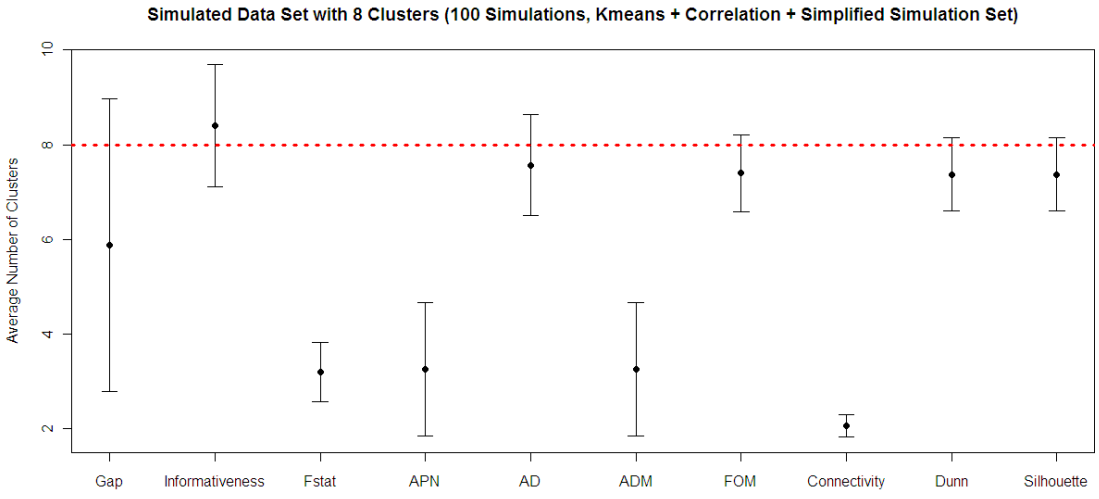
K=6	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	6.41	7.23	2.00	2.00	2.00	4.03	2.00	6.70	2.00	6.41
Standard Deviation	1.05	1.75	0	0	0	0.181	0	0.965	0	0.847
Range	[2,10]	[4,10]	[2,2]	[2,2]	[2,2]	[4,5]	[2,2]	[4,10]	[2,2]	[4,10]

Table 3: Performance of 10 Metrics on the 8-Cluster Data Sets

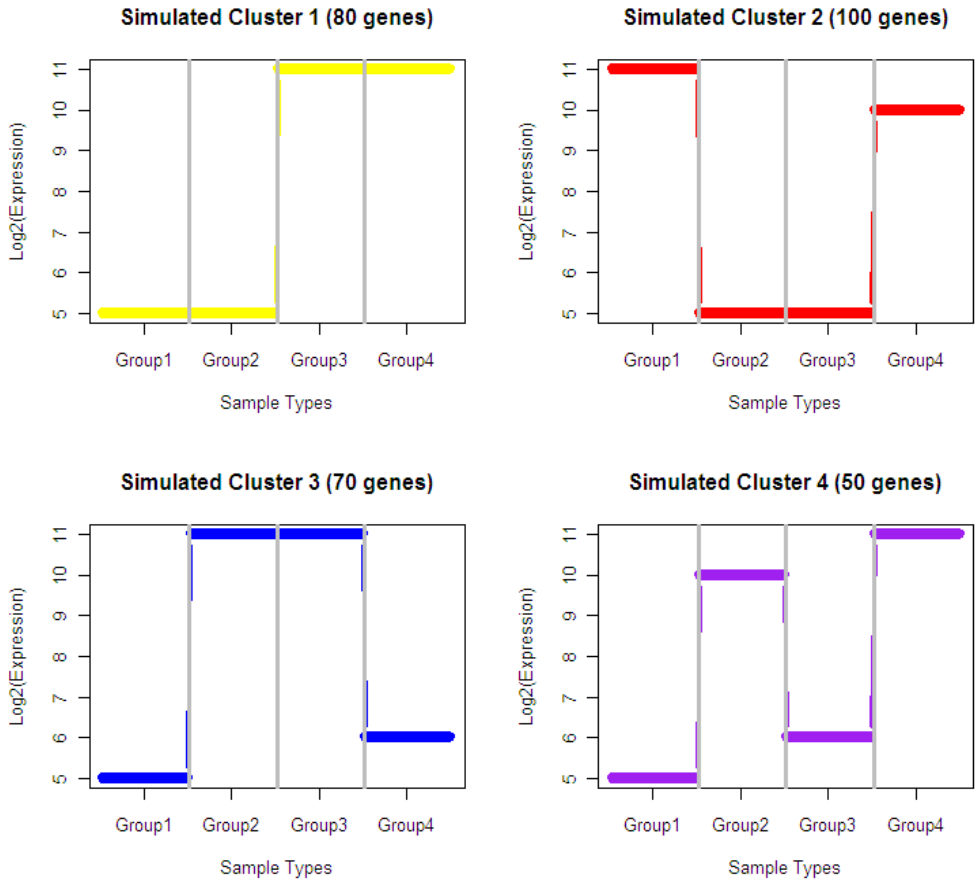
K=8	Gap Statistic	Informativeness Metric	F statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	5.89	8.4	3.2	2.06	7.37	7.37	3.26	7.57	3.26	7.40
Standard Deviation	3.09	1.29	0.632	0.236	0.770	0.770	1.40	1.07	1.40	0.812
Range	[3,12]	[6,12]	[2,5]	[2,3]	[6,8]	[6,8]	[2,8]	[6,10]	[2,8]	[6,9]


Simulated Data Set with 4 Clusters (100 Simulations, Kmeans + Correlation + Simplified Simulation Set)
Average Number of Clusters
2
3
4
5
6
7
8
Gap
Informativeness
Fstat
APN
AD
ADM
FOM
Connectivity
Dunn
Silhouette

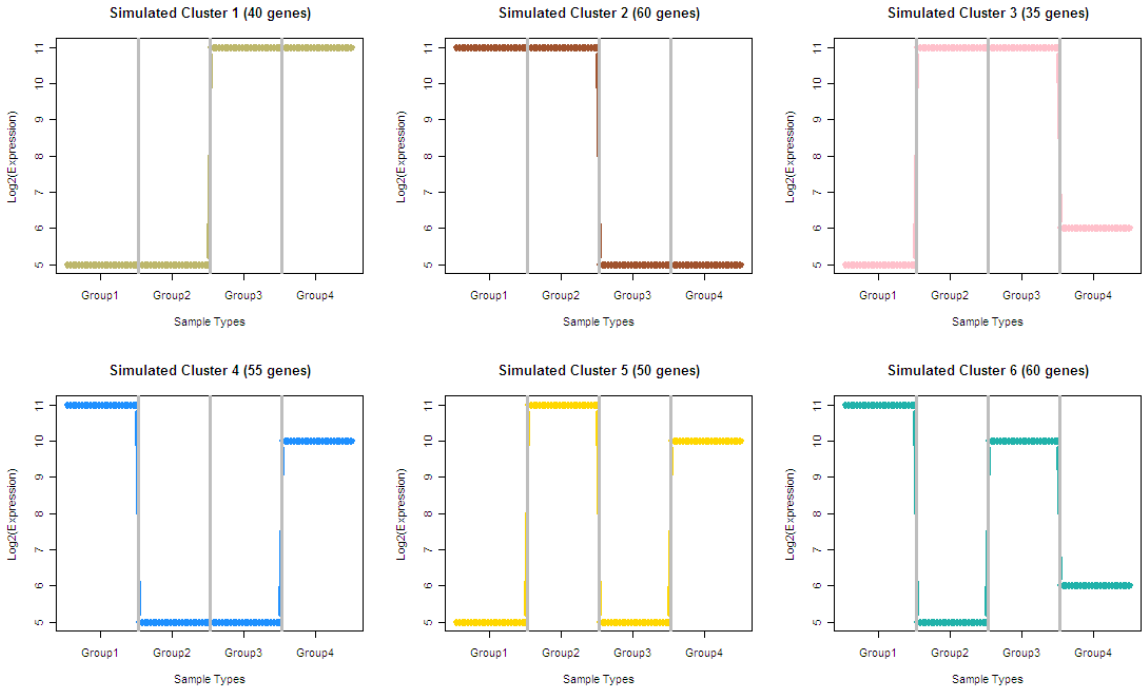

Simulated Data Set with 6 Clusters (100 Simulations, Kmeans + Correlation + Simplified Simulation Set)
Average Number of Clusters
2
3
4
5
6
7
8
9
Gap
Informativeness
Fstat
APN
AD
ADM
FOM
Connectivity
Dunn
Silhouette


Simulated Data Set with 8 Clusters (100 Simulations, Kmeans + Correlation + Simplified Simulation Set)
Average Number of Clusters
2
4
6
8
10
Gap
Informativeness
Fstat
APN
AD
ADM
FOM
Connectivity
Dunn
Silhouette

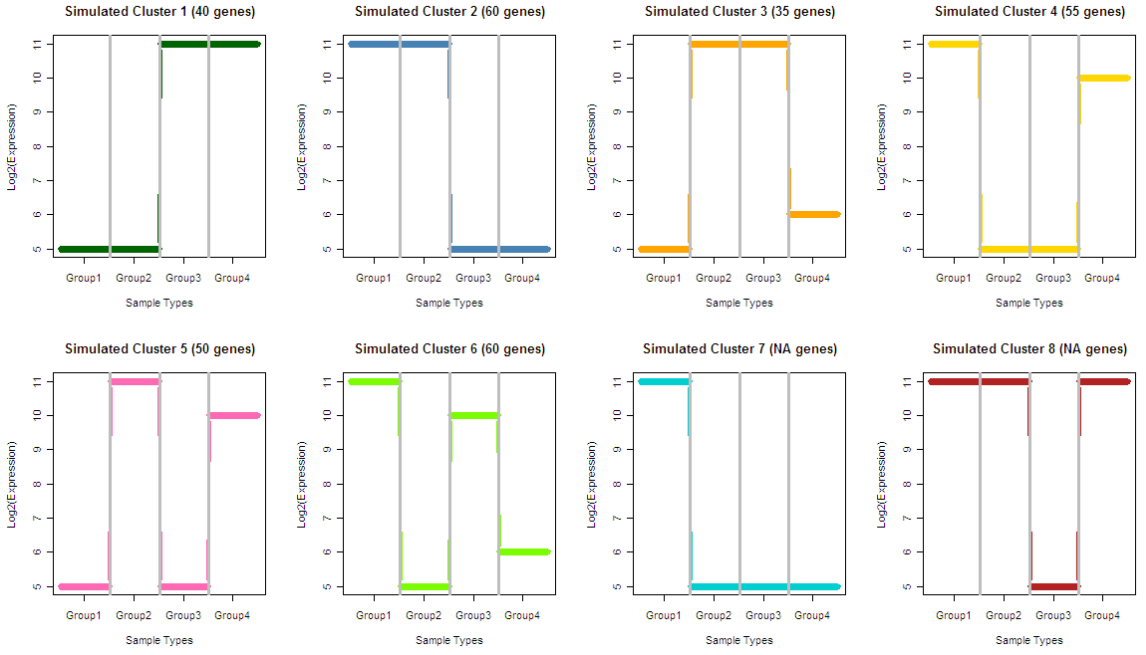
4-Cluster Data Set Simplified Reference Clusters



6-Cluster Data Set Simplified Reference Clusters

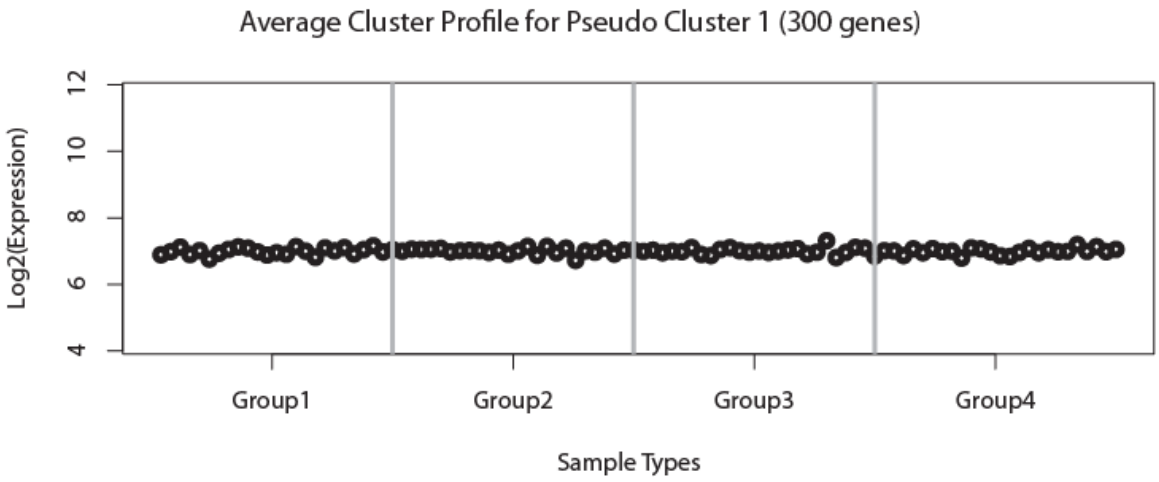


8-Cluster Data Set Simplified Reference Clusters

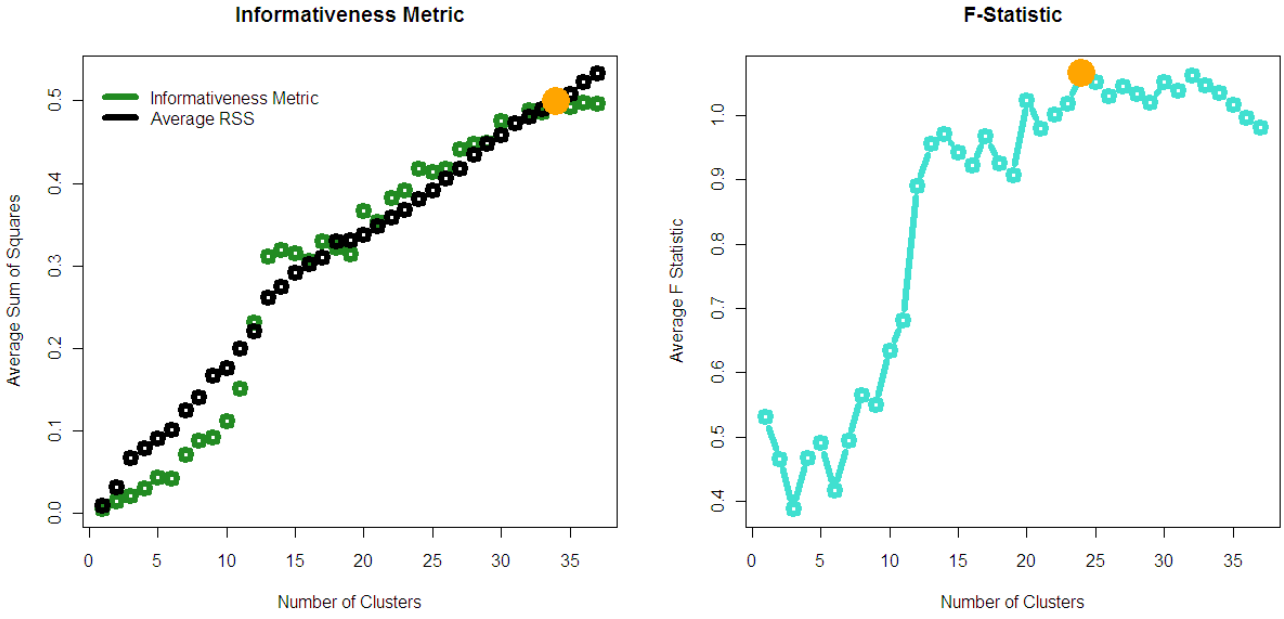


Supplementary Section: Inferring the Optimal Number of Clusters in the Absence of Group Structure

We simulated expression data for 300 pseudo-genes over 100 samples using a Normal distribution with mean parameter 7 and standard deviation 1.5. Our data shows a complete absence of group structure and all genes follow the same flat average profile.



The values over which the number of clusters was tested ranged from 1 to 33 clusters. The upper limit of this interval was determined by specifying the maximum number of clusters that gave rise to clusters with at least five genes in any given cluster.



The informativeness metric and average RSS statistic gave similar values across the entire test interval. The proximity of this curve with the average RSS values is a direct result of the absence of clustering structure in this data set. The maximum value of the informativeness metric is at the upper limit of the interval, however because the informativeness metric and average RSS values are overlapping, it is clear that the optimal number of clusters is not actually 34 but instead should be 1. The F-based

statistic instead is maximized at 24 clusters however from this curve alone, it is impossible to distinguish between a situation where there are genuinely 14 clusters versus no structure present.

Note the other metrics featured in this paper do not have the ability to test whether the optimal number of cluster is equal to one, but instead assume the existence of clustering structure and test the data for two or more clusters. We tested the estimates resulting from all ten metrics for 10 simulated data sets, all simulated under the same parameters described above. Interestingly, these metrics return estimates from either end of the testable spectrum – APN, ADM and the connectivity metrics estimate the existence of 2 clusters for all 10 data sets, whereas the Dunn Index, Silhouette Width, AD and FOM return estimates closer to the upper limits used.

Table 1: Performance of 10 Metrics on the One-Cluster Data Sets

K = 1	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM
Average	2.0	33.1	33.1	2.0	33.1	2.0	20.0
Standard Deviation	0	3.00	3.00	0	3.00	0	13.0
Range	[2, 2]	[29, 38]	[29, 38]	[2, 2]	[29, 38]	[2, 2]	[4, 37]

K = 1	Gap Statistic	Informativeness Metric	F statistic
Average	16.3	32.1	18.8
Standard Deviation	15.3	3.18	9.55
Range	[2, 37]	[28, 38]	[1, 33]

Supplementary Section: Testing Performance on the Müller Data Set

We applied a quality control filter to the original Müller data set, where probes with a detection score ≥ 0.99 in at least 75% of samples in the same cell type were retained, resulting in 9,956 probes that were retained for further analysis.

We fitted a LIMMA model to test for the significance of genes having altered expression across the four cell types. After adjusting the P-values with a Benjamini-Hochberg correction method, a considerable number of genes were significant for a change over any of the four cell types; a P-value threshold of 10^{-25} yielded a set of 1741 probes.

Using complete-linkage agglomerative hierarchical clustering with a Pearson correlation metric, the upper limit in the interval to test our metrics over was 27 (the maximum number of clusters to yield clusters with at least five genes).

The following results were obtained for the ten metrics:

	Compactness Metrics				Stability Metrics				F-statistic	Informativeness Metric
	Gap Statistic	Connectivity	Dunn Index	Silhouette Width	APN	AD	ADM	FOM		
Number of Clusters	2	2	3	2	2	27	2	27	2	6

The informativeness metric estimated that six clusters was the most appropriate number of clusters for this data set; Figure 1 illustrates how these six clusters capture distinct expression profiles that distinguish between all cell types. Six of the ten metrics estimated that two clusters was the most appropriate number of clusters. The two clusters resulting from this cluster analysis distinguish between the neural progenitor group and the other remaining cell types only (see Figure 2). While we cannot say what the exact number of true clusters describe this experimental data set, clearly the information arising from two clusters in Figure 2 is much more limited than what can be derived from the six clusters in Figure 1.

Figure 1: Average cluster expression profiles with the number of clusters prescribed by the informativeness metric.

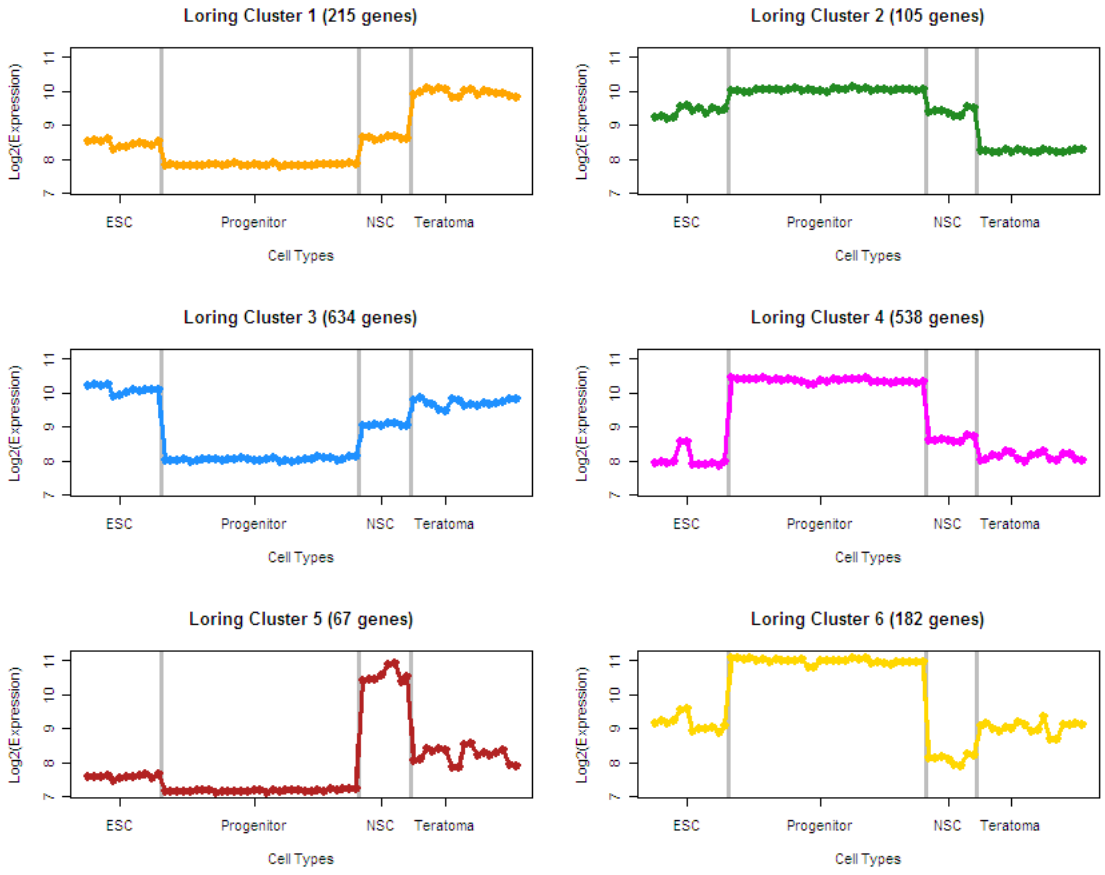
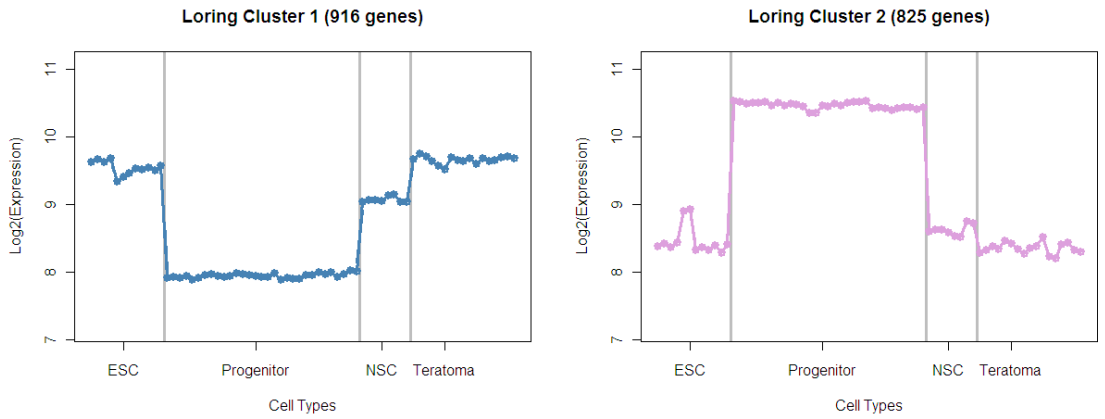
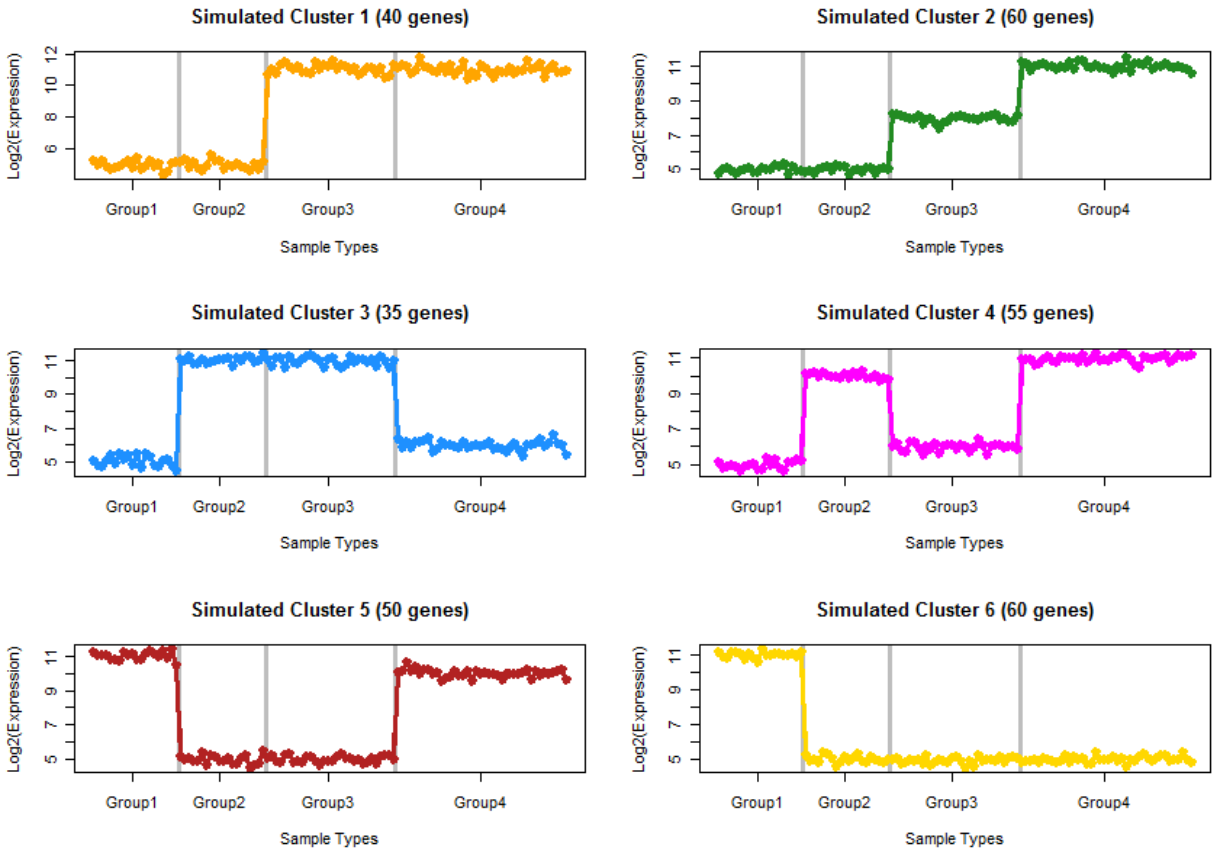


Figure 2: Average cluster expression profiles with the number of clusters prescribed by the Gap statistic, connectivity, Silhouette width, APN, ADM and F-statistics.



Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data where the Sample Sizes for Each Phenotypic Group is Different

We simulated gene expression data for 300 genes and 110 samples, corresponding to four phenotypic groups with 20, 20, 30 and 40 replicates per group. Using different mean values, we simulated six clusters in this data set under a Normal distribution, using a fixed standard deviation value of 1.5 throughout.

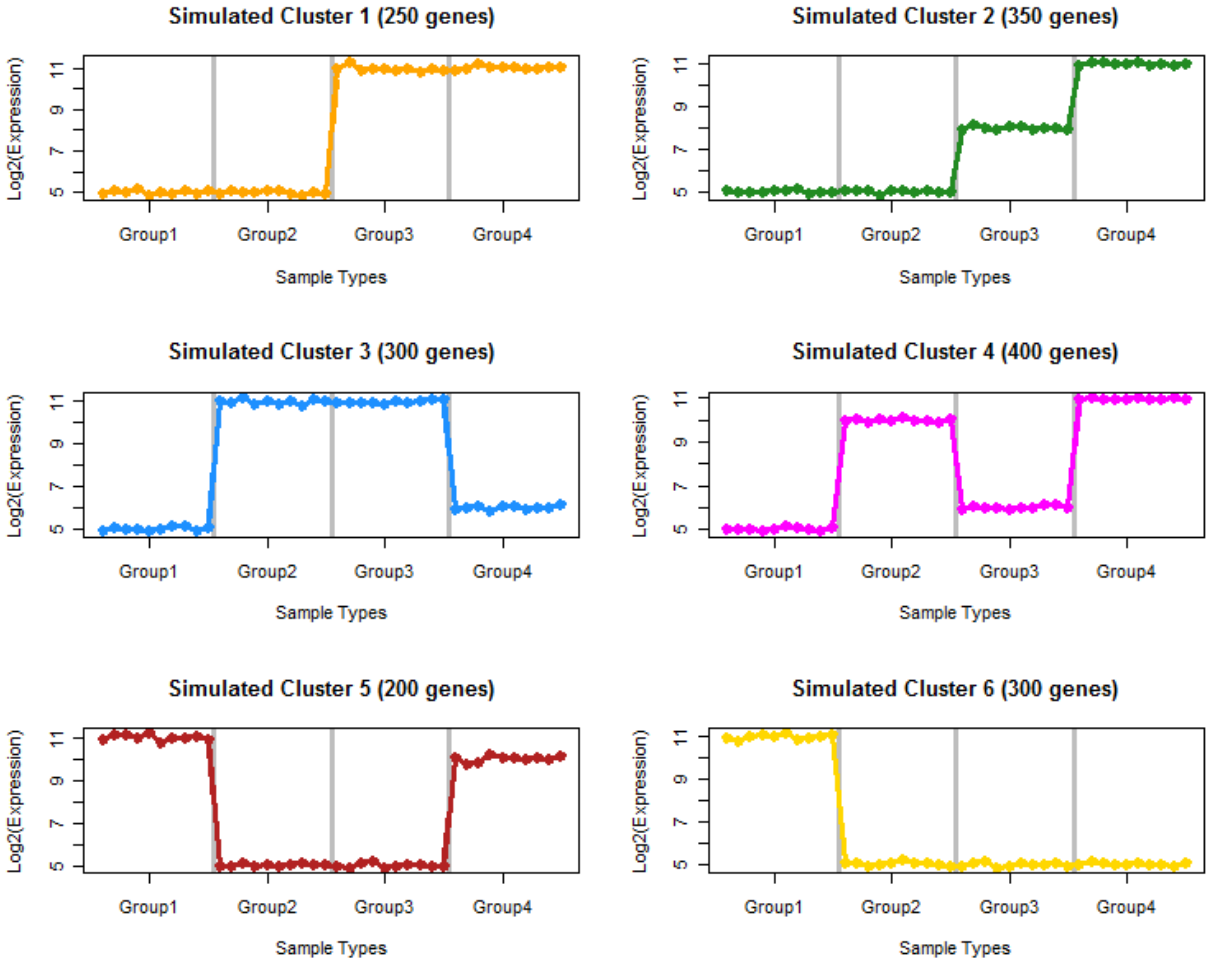


Using complete linkage agglomerative hierarchical clustering with Pearson correlation, the maximum number of clusters to split the genes into clusters that had at least five genes, was 9. We used the following metrics to test the most appropriate number of clusters between the interval of 2 to 9. Under these simulation parameters of unequal sample sizes, both the informativeness metric and the gap statistic were the only metrics able to recover the correct number of clusters in which this data set was simulated under, as illustrated in the table below.

	Compactness Metrics				Stability Metrics					
Number of Simulated Clusters	**Gap Statistic**	**Connectivity**	**Dunn Index**	**Silhouette Width**	**APN**	**AD**	**ADM**	**FOM**	**Modified F statistic**	**Informativeness Metric**
6	6	2	5	4	2	9	2	9	2	6

Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data for a Large Number of Genes with a Small Number of Samples

We simulated gene expression data for 1800 genes and 40 samples, corresponding to four phenotypic groups with 10 replicates per group. Using different mean values, we simulated six clusters in this data set under a Normal distribution, using a fixed standard deviation value of 1.5 throughout.

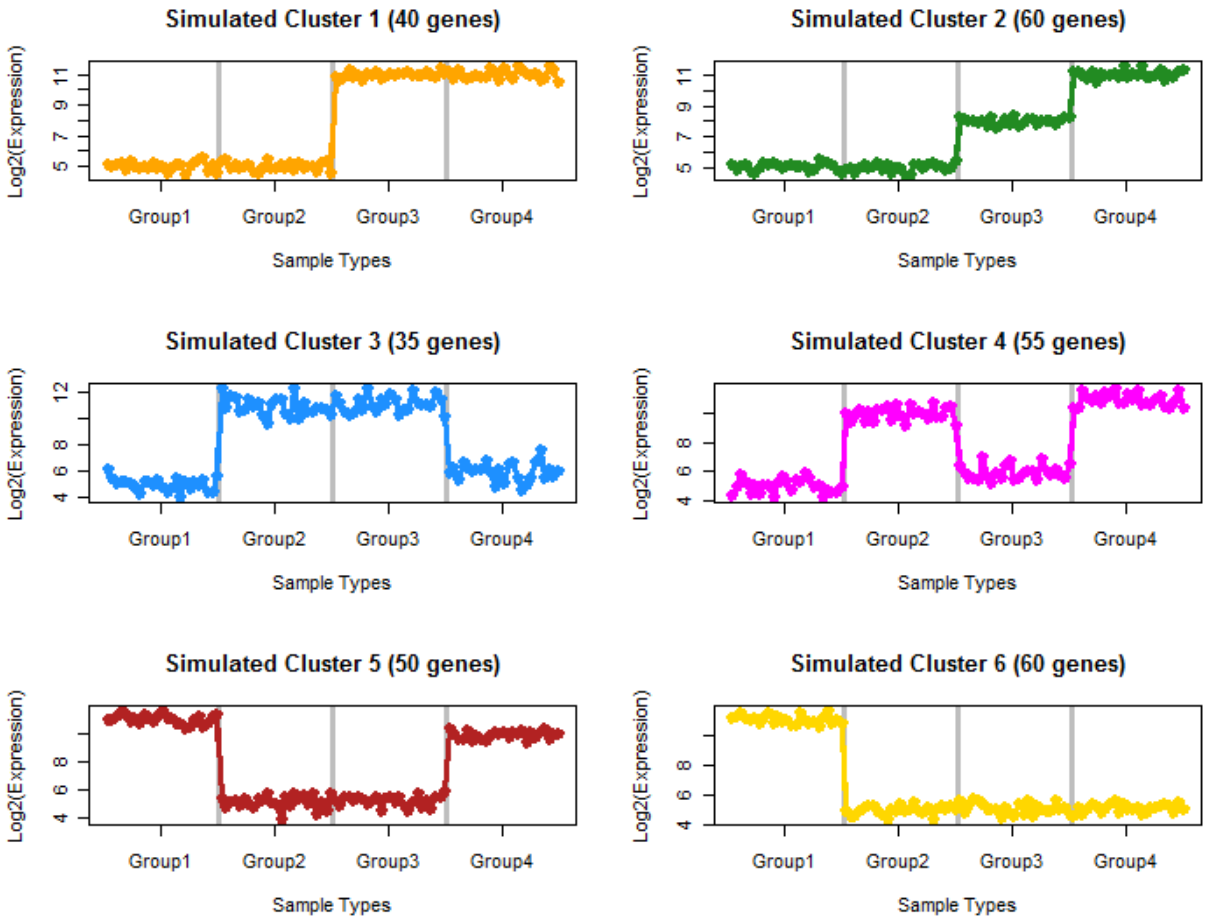


Using complete linkage agglomerative hierarchical clustering with Pearson correlation, the maximum number of clusters to split the genes into clusters that had at least five genes, was 7. We used the following metrics to test the most appropriate number of clusters between the interval of 2 to 7. Under these simulation parameters of large genes to a small number of samples, both the informativeness metric and the gap statistic were the only metrics able to recover the correct number of clusters in which this data set was simulated under, as illustrated in the table below.

	Compactness Metrics				Stability Metrics					
Number of Simulated Clusters	**Gap Statistic**	**Connectivity**	**Dunn Index**	**Silhouette Width**	**APN**	**AD**	**ADM**	**FOM**	**Modified F statistic**	**Informativeness Metric**
6	6	2	4	4	2	7	2	7	2	6

Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data where Each Cluster is Simulated Under a Different Variance Parameter

We simulated gene expression data for 300 genes and 100 samples, corresponding to four phenotypic groups with 25 replicates per group. Using different mean values, we simulated six clusters in this data set under a Normal distribution, where each cluster was simulated under a different standard deviation value. For the six clusters simulated, we used standard deviations of 1.5, 2.0, 3.5, 4.0, 3.0, 2.5.

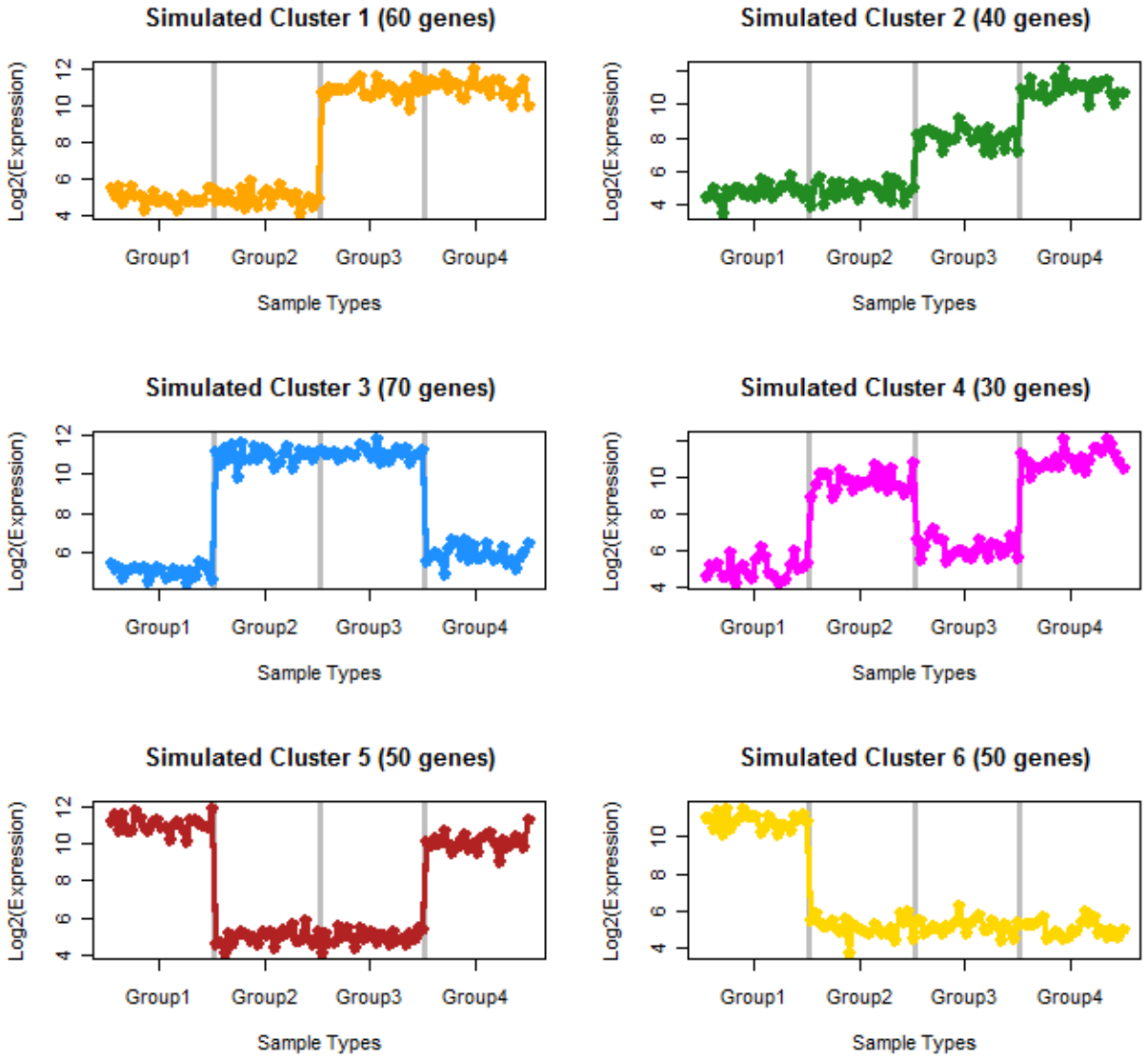


Using complete linkage agglomerative hierarchical clustering with Pearson correlation, the maximum number of clusters to split the genes into clusters that had at least five genes, was 13. We used the following metrics to test the most appropriate number of clusters between the interval of 2 to 13. Under these simulation parameters of a different variance parameter for each cluster, both the informativeness metric and the gap statistic were the only metrics able to recover the correct number of clusters in which this data set was simulated under, as illustrated in the table below.

	Compactness Metrics				Stability Metrics					
Number of Simulated Clusters	**Gap Statistic**	**Connectivity**	**Dunn Index**	**Silhouette Width**	**APN**	**AD**	**ADM**	**FOM**	**Modified F statistic**	**Informativeness Metric**
6	6	2	3	4	2	13	2	13	4	6

Supplementary Section: Testing Performance of Metrics on a Simulated Gene Expression Data where Genes within Each Cluster Have Different Variances

We simulated gene expression data for 300 genes and 100 samples, corresponding to four phenotypic groups with 25 replicates per group. Using different mean values, we simulated six clusters in this data set under a Normal distribution, where each cluster was simulated so that genes within a cluster had different variances. We implemented this simulation design by forcing half of the genes in each cluster to have a standard deviation of 1.5 and the other half to have a higher standard deviation of 4.



Using complete linkage agglomerative hierarchical clustering with Pearson correlation, the maximum number of clusters to split the genes into clusters that had at least five genes, was 10. We used the following metrics to test the most appropriate number of clusters between the interval of 2 to 10. Under these simulation parameters of genes adopting different variance parameters within a single cluster, none of the metrics were successful in recovering the correct number of clusters in which this data set was simulated under, as illustrated in the table below. However, of all the estimates, the informativeness metric was the closest to the real value, at 5 clusters.

	Compactness Metrics				Stability Metrics					
Number of Simulated Clusters	**Gap Statistic**	**Connectivity**	**Dunn Index**	**Silhouette Width**	**APN**	**AD**	**ADM**	**FOM**	**Modified F statistic**	**Informativeness Metric**
6	3	2	3	4	2	10	2	10	2	5