Single-cell RNA Expression Clustering with the EM and Hierarchical Algorithms

Robert Tumasian III, Ling Tuo, and Qetsiyah Wang

April 7, 2021

Introduction

Single-cell RNA-sequencing (scRNA-seq) is a novel biological technique used to measure gene expression levels in individual cells. This innovation has enabled us to gain a more thorough understanding about the underlying functionality of diverse cell types. An area of particular interest in scRNA-seq is cell heterogeneity; that is, cells of the same type can have vastly different gene expression levels. This heterogeneity can be due to several factors, including cell age, environmental conditions, and the existence of cell subtypes. Clustering analysis can be used to gain knowledge about these subtypes.

Using a data set containing gene expression levels from cells from breast cancer tumors, the goals of this study are to:

(1) Identify how many principal components are able to explain sufficient variability in the data using the `prcomp` function in `R`
(2) Build a Gaussian-Mixture model with the principal components from (1) using the expectation-maximization (EM) algorithm to assess cluster classification accuracy
(3) Determine which genes are most important for differentiating the clusters
(4) Evaluate the suitability of this approach
(5) Consider the hierarchical clustering algorithm
(6) Summarize and compare the results of the two algorithms

Methods

Data Description

This study utilizes gene expression level data from breast cancer tumors, containing 557 genes from 716 cells. Additionally, about 80.5% of the gene expression data is zeros.

Principal Component Analysis (PCA)

To conduct PCA, the gene expression data was scaled by dividing each entry by its column sum. Scaling is essential for PCA, since it is a variance-maximizing procedure. General normalization procedures are not able to be used due to excess zeros in the data. To select an appropriate number of principal components (PCs), we considered three approaches: (1) the cumulative proportion of explained variance criterion [number of PCs needed to capture 80% of the total variance], (2) Kaiser's rule [retaining all PCs that have a variance larger than the average variance], and (3) Scree plots [selecting the number of PCs at which the Scree curve begins to flatten out].

K-means Clustering

The goal of the K-means algorithm is to iteratively partition the data set into K pre-specified, non-overlapping clusters, in which every data point only belongs to one cluster. The algorithm aims to make the intra-cluster data points as similar as possible but keep the clusters far apart for differentiation. Each data point is assigned to a cluster based on its (Euclidean) distance from each cluster's centroid. The point is then assigned to the cluster that is closest (smallest distance to centroid). The algorithm stops when all centroid values do not substnatially change (data point assignment remains the same).

The K-means algorithm is a variant of the expectation-maximization (EM) algorithm, which is discussed below. While K-means uses hard assignment based on convergence and uses the L2 norm for optimization, the EM algorithm uses a soft, probabilistic assignment and depends on the expectation of a point belonging to a particular cluster rather than on the L2 norm.

Gaussian-Mixture Model (GMM)

GMMs are commonly used to investigate normally distributed subpopulations within a larger population. These models are unsupervised, since the assignment of the data points to each subpopulation is unknown. In addition, GMMs are often fairly efficient with large data sets due to their similar computational and theoretical properties to general Gaussian models. A GMM consists of multiple $k = 1, ..., K$ Gaussians, where K denotes the number of clusters of our data set. Each Gaussian has a center (μ_k), a covariance (Σ_k) defining its width, and a mixture probability (π_k) that determines the size of the Gaussian function, where $\sum_{k=1}^{K} \pi_k = 1$. For K clusters, the GMM can be expressed as:

$$N(x|\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_k|}} exp(-0.5(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i))$$

.

When the number of clusters is known, the expectation-maximization (EM) algorithm, explained below, can be used to estimate the parameters of a GMM.

Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative, unsupervised technique used to overcome the challenges of the standard maximum likelihood estimation procedure when dealing with mixture models that have many parameters. The algorithm can only be used if closed form expressions are available for updating the model parameters, and it has the helpful property that the maximum likelihood of the data continues to increase at each iteration; thus, a local maximum or saddle point is guaranteed. The EM algorithm has three steps. First, for initialization, K different random data points are selected to serve as the cluster means and all cluster variances are set to the overall sample variance. Then, in the expectation step (E-step), the expectations of the cluster assignments for each data point, given all the model parameters, are calculated. Lastly, in the maximization step (M-step), all the expectations from the E-step are maximized with respect to the model parameters and all parameters are updated. This process occurs iteratively until convergence of all parameter estimates is achieved.

Hierarchical Clustering Algorithm

There are also other techniques which can be used for clustering analysis. Hierarchical clustering, an alternative method of partitional clustering, starts form dividing the whole data set into two homogeneous clusters and then partitioning the rest of data into two new homogeneous clusters for each existing cluster. The algorithm will repeat iteratively until the clusters produce identical items. There are two types of hierarchical clustering:

agglomerative and divisive. Agglomerative clustering starts from the root, that each observation is regarded as one single cluster and then merged with other similar observations. In the end, all single clusters are merged into one jumbo cluster. Divisive is works oppositely. Unlike partitional clustering, such as K-means, hierarchical clustering does not require a pre-specified K value; however, the analysis quality largely depends on the distance measure and linkage function chosen for clustering. Distance measures include Euclidean, maximum, manhattan, or correlation for measuring dissimilarities between each observation. Five linkage functions are also available (single, complete, centroid, average, and Ward's minimum variance) for measuring inter-cluster distance.

In this report, we will focus on divisive hierarchical clustering and its implementation on exploring the gene expression for each cell. The original microarray data was $log2$-transformed and scaled up by 0.25 to address zero values. Spearman correlation distance and complete linkage were selected.

Silhouette Score Method

For obtaining the optimal K to initiate clustering, the silhouette score method was employed. The silhouette score is a metric for testing the goodness of clustering for the target sample by measuring the average intra-cluster distance (a) and the average inter-cluster distance (b) for each observation.

$$Silhouette\ score = \frac{(b_i - a_i)}{max(a_i, b_i)},$$

where b_i is the distance between observation i and its nearest cluster which i does not belong to. The silhouette score is between -1 and 1, where values approaching to 1 indicate higher goodness of clustering.

Results

For determining the number of PCs that were sufficient to capture the variability of the gene expression data, we obtain 7, 34, and 6 PCs using the cumulative proportion of explained variance criterion at 80% (Table 1), Kaiser's rule, and a Scree plot (Figure 1), respectively. We proceeded with the 6 PCs from the Scree plot.

After estimating a multivariate GMM with the PCs found in Part 1 using the EM algorithm, we used the silhouette scoring method to identify the optimal number of clusters. The largest silhouette score was achieved using K=4 clusters (Figure 2). Clustering results using 6 PCs, the optimal K=4 clusters, and the GMM with the EM algorithm are provided in Figure 3. We found that most cells were assigned to cluster 2 (in red) and only one cell was assigned to cluster 1 (in black) (Figures 3 and 4). This is likely due to the cell being an outlier or having gene expression patterns very different from those of other cells (Figure 4). Removing this cell would likely result in the optimal number of clusters being three instead of four.

Next, the top 10 and 50 genes with the most expression across all cells were extracted and compared between the four clusters. Among the top 10 genes, there was some overlap in the genes most expressed in the clusters (e.g. Rn45s and Mgp) (Table 2). There was more considerable overlap among the top 50 most expressed genes across all cells (Figure 5). We postulate that the most expressed genes that are unique to each cluster are the most important genes for differentiating the clusters. For clusters 1-4, there are 20 (Lum, Col8a1, Dcn, Mfap5, Mgst1, Ndufa4, Pi16, Tnc, Lox, Glipr1, Ccl2, Ccnb2, Prelp, Csn3, Sod3, Tspan7, Trf, Lrrc15, Sfrp2, and Cpxm2), 1 (Ly6e), 4 (Rgs16, Tinagl1, H2.M9, and Apoa1bp), and 2 (Thbs2 and Nupr1) such genes, respectively, that are most helpful for differentiation.

Employing the GMM with PCA for clustering may not be a suitable approach, since it is sensitive to outliers and does not have a built-in method to account for data with excess zeros. Perhaps using a different clustering or distance method that is less sensitive to outliers, or implementing a model that can better handle zero-inflated data, would yield more reliable results.

Finally, we implemented hierarchical clustering to compare its results to the GMM with PCA. The hierarchical dendrogram, shown in Figure 6, shows the trend of divisive clustering. Horizontal lines represent the

dissimilarity or distance between the cells and X_{height} at merge points at which two horizontal joined lines indicate the magnitude of that dissimilarity. In the dendrogram, correlation distance was utilized, so height would be between 0-2. The higher dissimilarity is, the further two individual cells will be. For instance, the most bottom blue cell would merge together with the top yellow line at a high distance of 0.9099, meaning that the certain gene expression would be largely different in these two cells. Due to no requirement of a pre-specified K, four groups were used for comparison between the hierarchical and K-means algorithms. Grouping is established by cutting the tree with one vertical line in the dendrogram. Four clusters are presented in different colors. Within the blue cluster, cells presented a high merge point at 0.7848 of 0.7848, indicating that there may be a cell outlier existing in blue cluster with consideration of the highest height point mentioned before. For the red cluster, the same conclusion could also be made from both the inter- and intra-cluster merge points of 0.8329 and 0.7601, respectively, that there would be a cell outlier present. Generally, hierarchical clustering is helpful for differentiating the gene expression signatures in different cells through calculating dissimilarities. However, for large dimensional data, hierarchical clustering would be too complicated for calculating the dissimilarities between individual clusters and the resulting dendrograms would be very difficult to visualize and assess.

Like K-Means clustering, hierarchical clustering is also able to incorporate feature selection performed by PCA analysis. For the dendrogram in Figure 7, we again utilized Spearman correlation distance and complete linkage, excluding the $log2$ scale for better comparison with K-means. Incorporating feature selection by PCA yielded a more visually understandable dendrogram. Four groups were also selected for this dendrogram, where cluster 1 is in red, cluster 2 is in green, cluster 3 is in blue, and cluster 4 is in purple. Notice the first merge point at a height of 2, meaning that either clusters 1 or 2 contained data points totally uncorrelated with either clusters 3 or 4. Considering each first intra-cluster merge point, excluding cluster 1, clusters 2, 3, and 4 all contained intra-outliers that were extremely far away from their other intra-cluster data points with height differences between their next merge points of 0.1143, 0.4572, and 0.5142, respectively. Comparing these results to the hierarchical clustering without PCA modification, the dissimilarity scale for each individual cell increased significantly, meaning that hierarhical clustering is largely impacted by potential missing data, especially for high-dimensional data.

Based on two clustograms (Hierarchical [left] vs. K-means [right]) in Figure 8, the most noticeable between the two algorithms is their difference in dealing with cell outliers. Hierarchical clustering tends to consider all data points into clustering, including outlier cell 314, largely decreasing the quality of clustering. In contrast, K-means clustering would divide outlier cell 314 into its own cluster (cluster 1), resulting in higher quality clustering. Besides plotting, we also can use silhouette scores to compare the two clustering methods.

From Table 3, for hierarchical clustering, clusters 3 and 4, which both included outliers, presented bad performance on clustering (negative silhouette scores). Generally, a higher silhouette score for K-means indicates that hierarchical clustering indeed presents more sensitivity to outliers and lower clustering quality. Above all, both the K-Means and hierarchical methods show advantages on different aspects of unsupervised data analysis. K-Means, by locating centroids for K clusters, shows higher efficiency in grouping homogeneous data. It is more flexible in dealing with outliers and able to produce much tighter clusters. However, the initial starting value for K strongly impacts the final clustering results, making it less useful with high-dimensional data. Due to random sampling used in locating centroids, the clustering process is hard to reproduce if no initial seed is set. Compared to K-Means, hierarchical clustering does not require an initial K value, decreasing analysis bias. And from the visualization aspect, hierarchical clustering is more informative, since dendograms can show apparent dissimilarity metrics for each individual studying object (e.g. cells), compared to overly condensed and unstructured plots of cluster points. Based on dendrograms, it is more efficient to determine how many clusters should be used. However, hierarchical clustering is also not suitable and performs even worse than K-Means for high-dimensional data because calculations of dissimilarity for each individual unit could be very computationally challenging. Its high sensitivity to outliers also decreases its clustering ability.

Appendix

Tables

Table 1: Number of PCs (7) needed to explain at least 80% of the total variability in the gene expression data (cumulative proportion of explained variance criterion at 80%).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	6.971	2.810	2.494	2.043	1.861	1.406	1.356
Proportion of Variance	0.528	0.086	0.068	0.045	0.038	0.022	0.020
Cumulative Proportion	0.528	0.614	0.682	0.727	0.765	0.786	0.806

Table 2: Top 10 genes with the most expression across all cells by cluster.

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Rn45s	Rn45s	Rgs5	Mgp
Mgp	Mgp	B2m	Malat1
Lars2	B2m	Rn45s	Col3a1
Col1a2	Cst3	Cst3	Fth1
Col3a1	Postn	Mgp	B2m
Olfml3	Crip1	Crip1	Anxa1
Lum	Malat1	Arhgdib	Cst3
Col8a1	Fth1	Col3a1	Crip1
Cyr61	Col3a1	Anxa1	Rn45s
Dcn	Anxa1	Higd1b	Hspa5

Table 3: Silhouette scoring to compare the hierarchical and K-means clustering algorithms.

Method	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Avg.Score**
Hierarchical	0.2755	0.1380	-0.0697	-0.0400	0.0759
K-means	0	0.4520	0.3082	0.1347	0.2237

Figures

Figure 1: Scree plot displaying the amount of variance explained by the first 10 PCs.

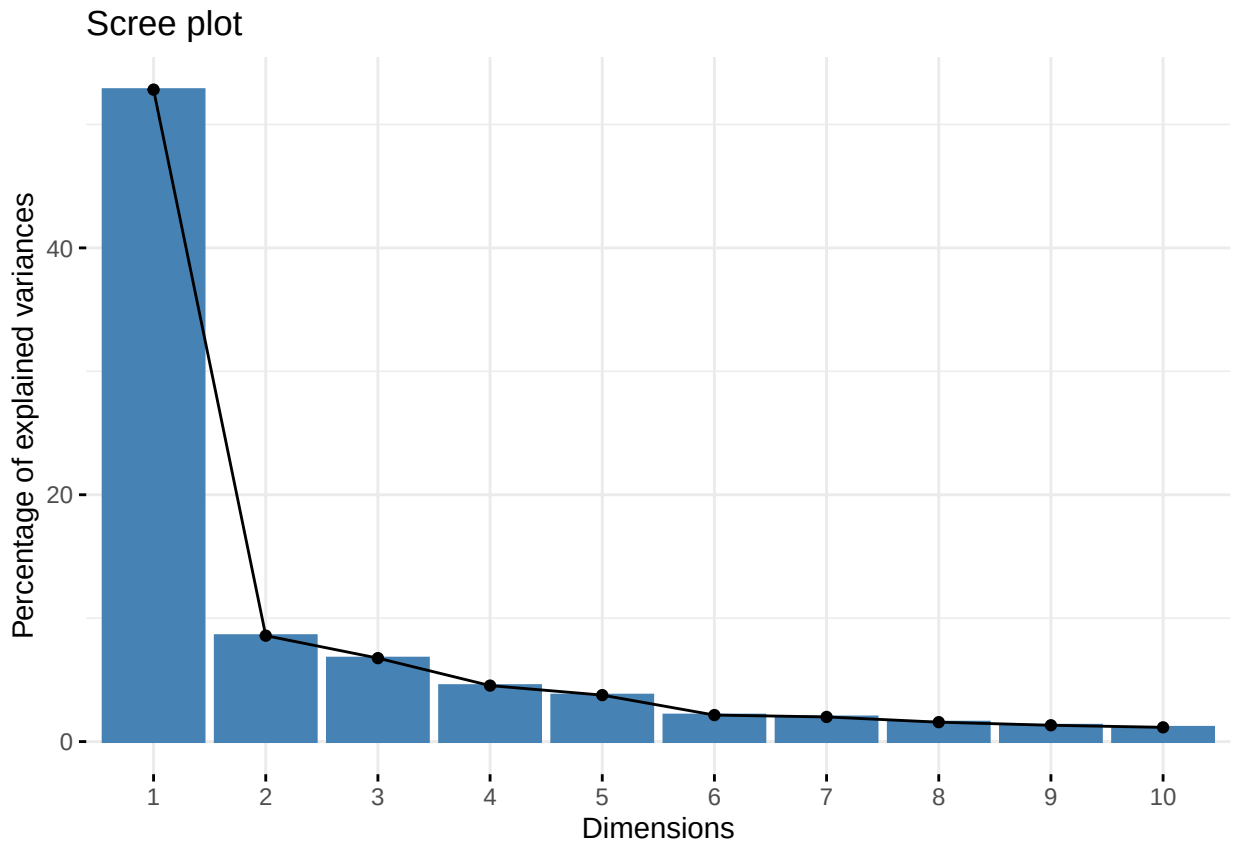

Figure 2: Silhouette score results using the 6 PCs identified by the Scree plot above.

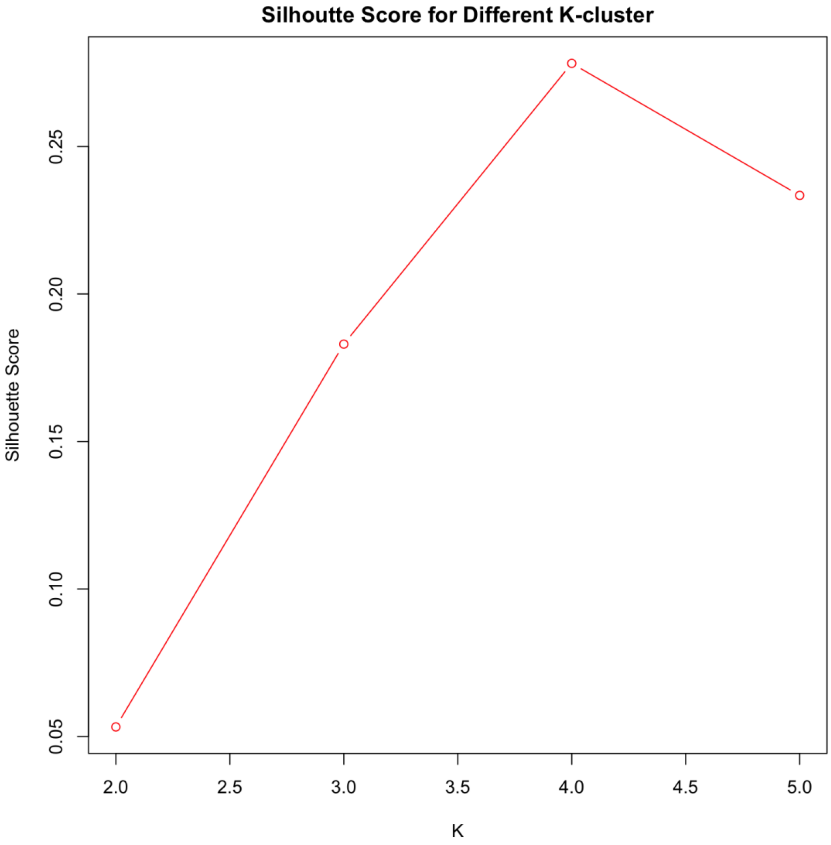

Figure 3: Clustering assignments for each cell using the optimal K=4 clusters and 6 PCs.

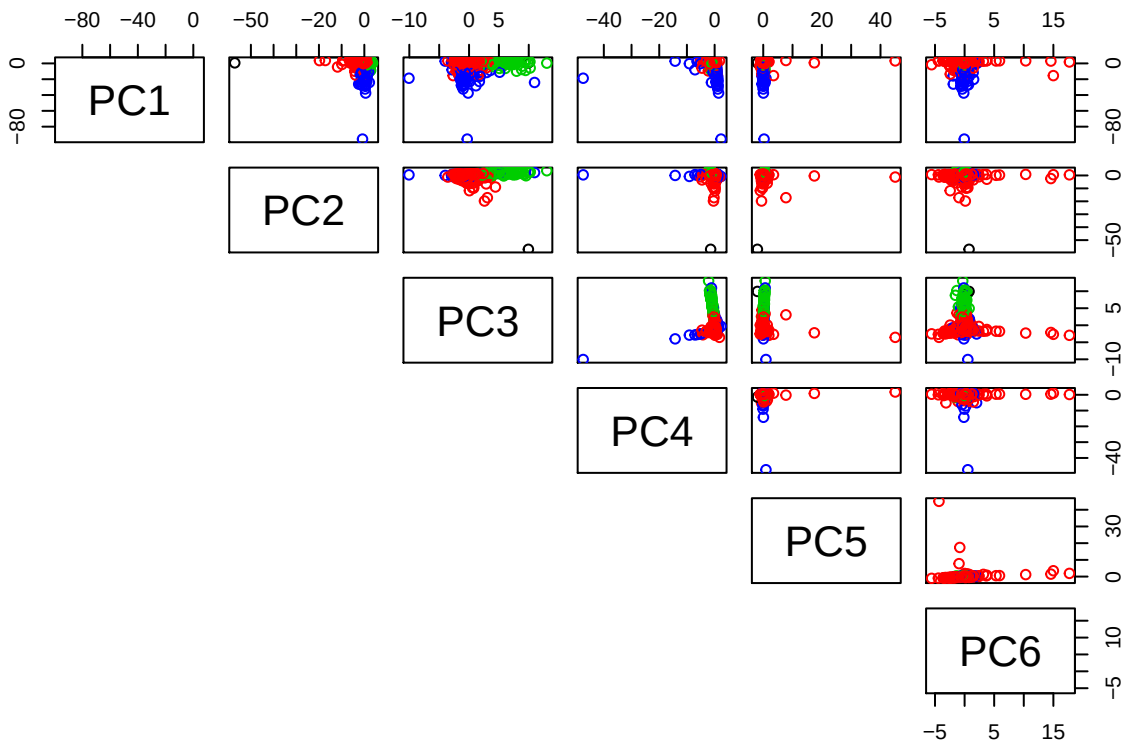

Figure 4: Cluster assignments for each cell.

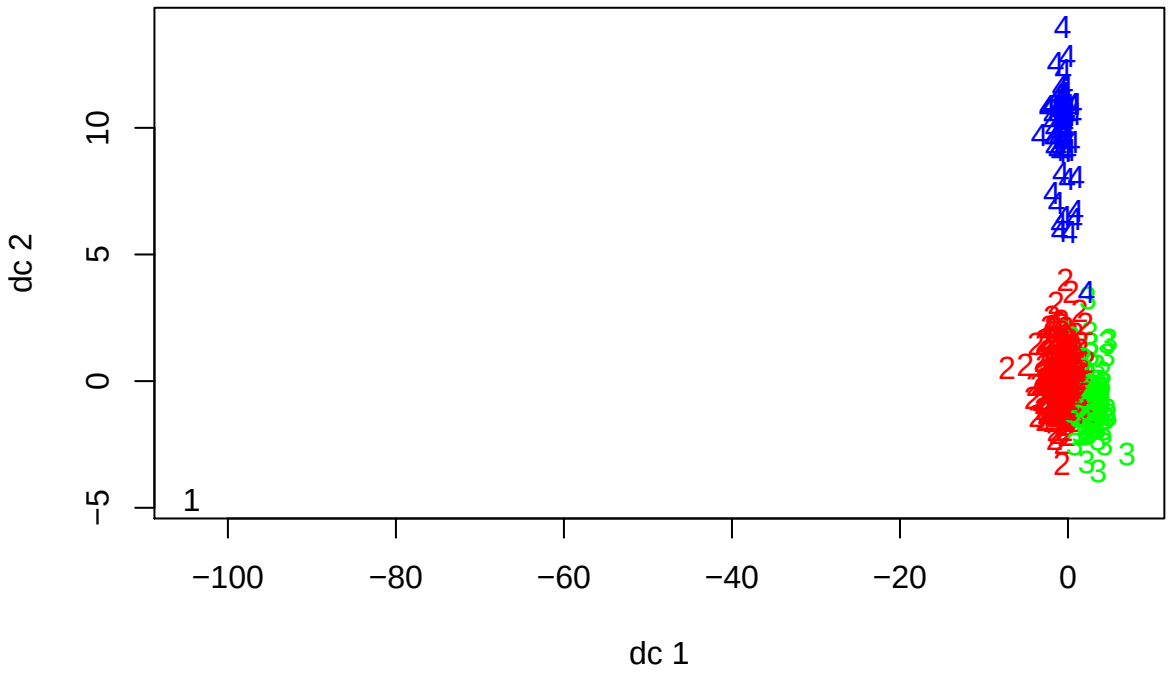

Figure 5: Venn diagram showing the overlap of the top 50 most expressed genes by cluster.

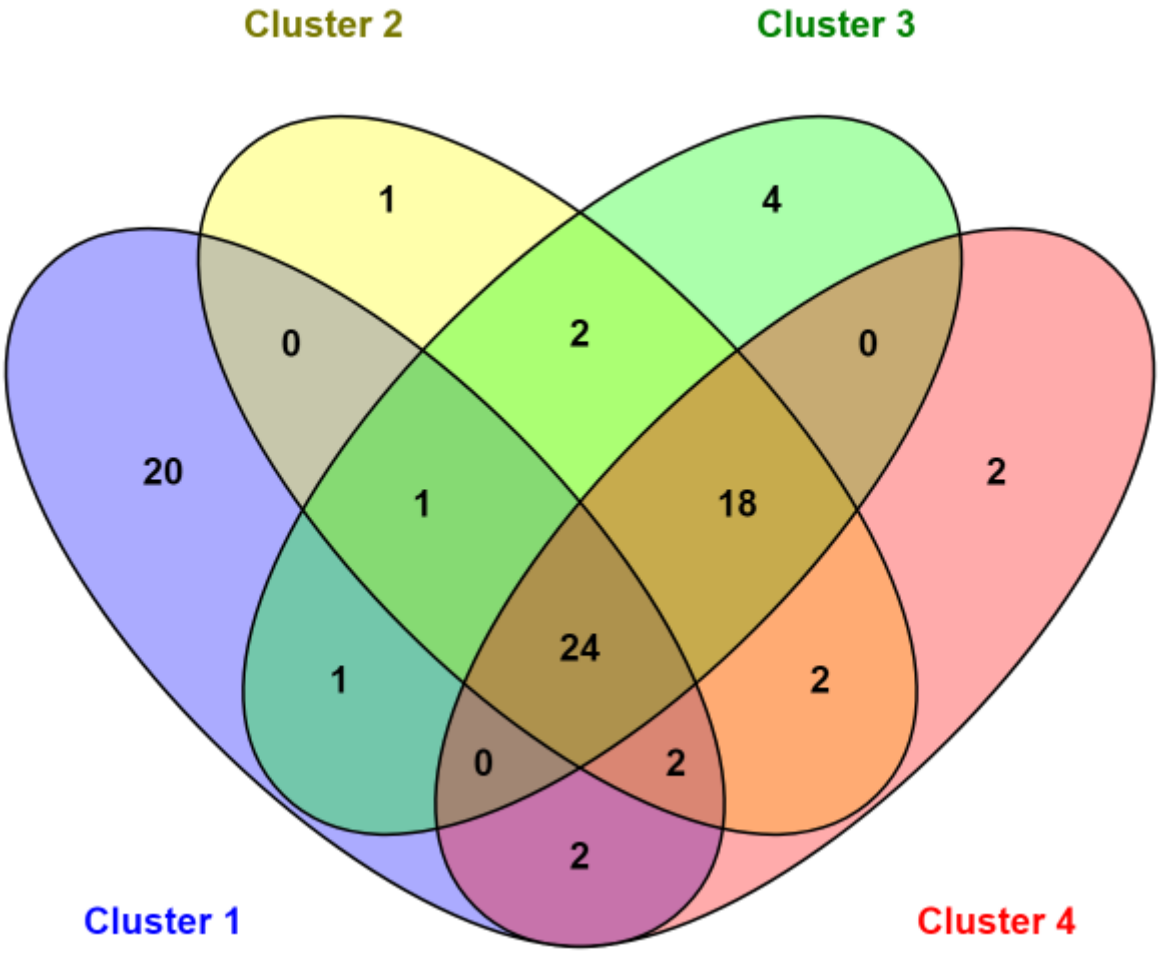

Figure 6: Hierarchical divisive clustering dendrogram displaying cluster assignments for each cell.

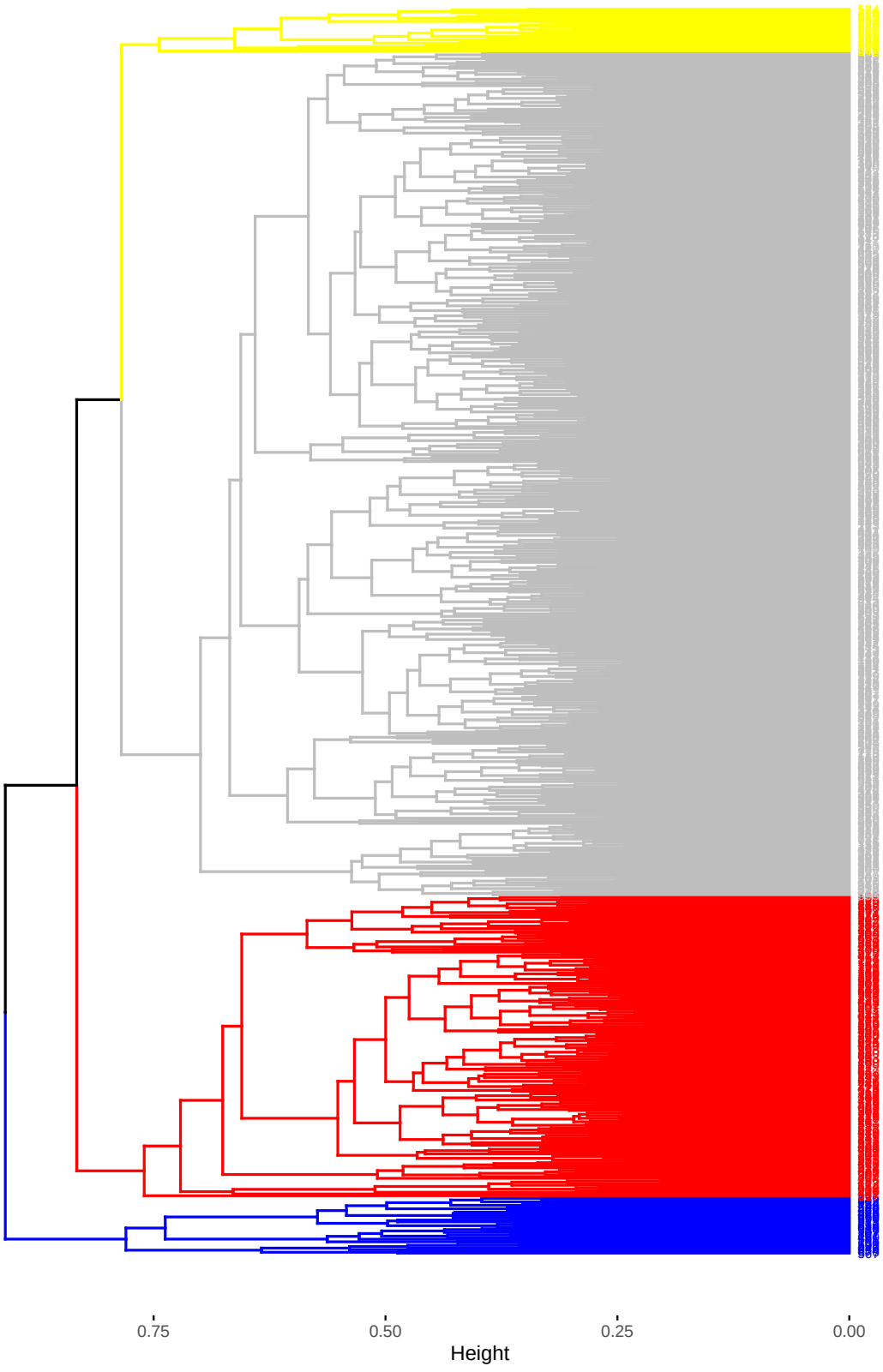

Figure 7: Hierarchical clustering dendrogram incorporating PCA.

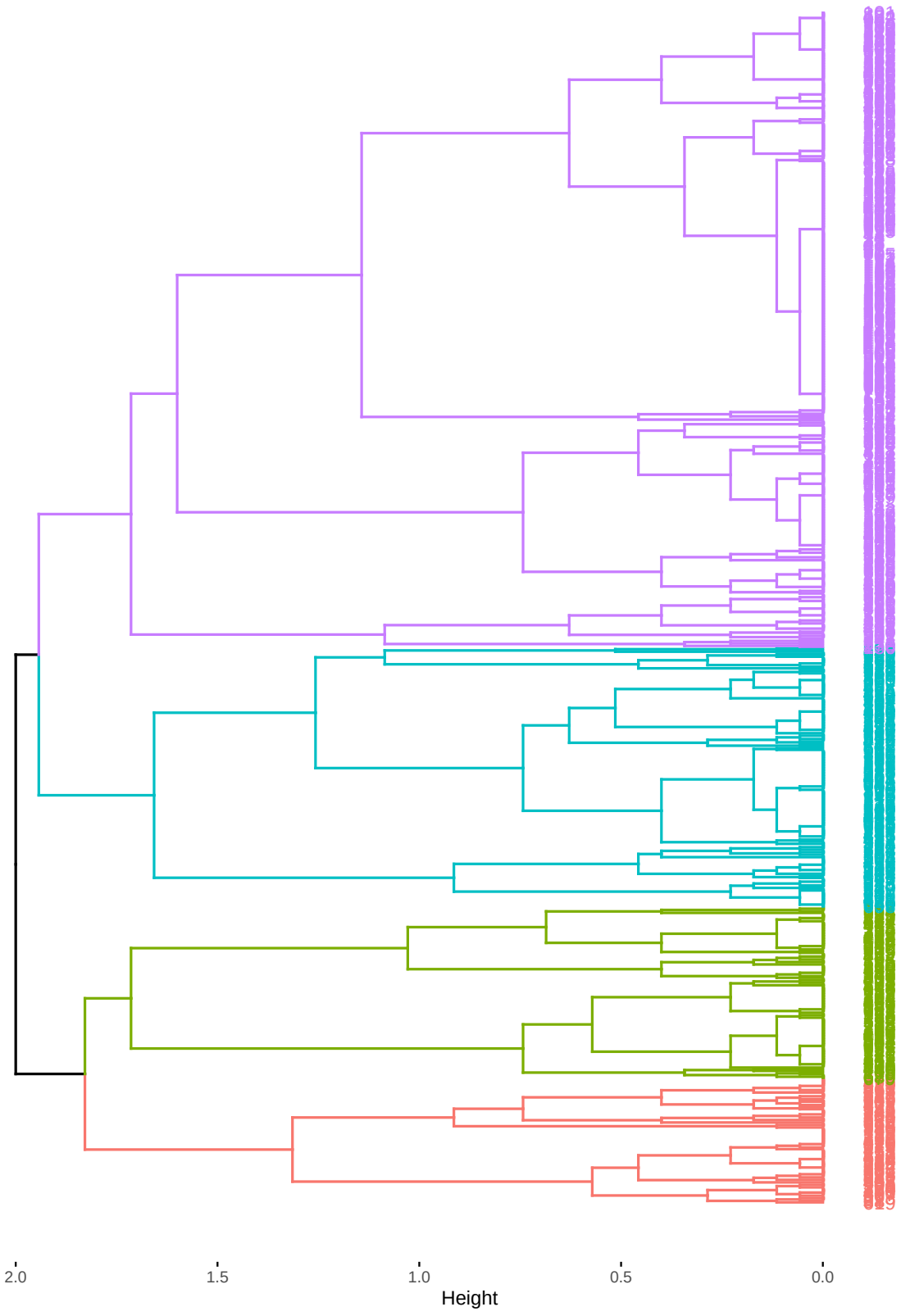

Figure 8: Clustograms comparing the hierarchical (left) and K-means (right) clustering algorithms.