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

Robert Tumasian III, Ling Tuo, and Qetsiyah Wang

Columbia University
Department of Biostatistics
Mailman School of Public Health

April 12, 2021

Outline

1. Background
2. Objectives
3. Methods
4. Results
5. Discussion

Background

- 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
- Cell heterogeneity is of particular interest; that is, cells of the same type can have vastly different gene expression levels
  - May be due to cell age, environmental conditions, or the existence of cell subtypes
  - Clustering analysis can be used to gain knowledge about these potential subtypes

Objectives

1. Identify how many principal components (PCs) are able to explain sufficient variability in the gene expression data
2. Build a Gaussian-mixture model (GMM) with the PCs from above using the expectation-maximization (EM) algorithm for clustering the cells
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: Principal Component Analysis (PCA)

- Gene expression data was scaled by dividing each entry by its column sum
  - Scaling is essential since PCA is a variance-maximizing procedure
  - General normalization was not used due to excess zeros in the data (80.5%)

- To select an appropriate number of PCs, we considered three different approaches:
  1. Cumulative proportion of the explained variance (number of PCs needed to capture 80% of the total variance)
  2. Kaiser's rule (retaining all PCs with a variance larger than the average variance
  3. Scree plots (selecting the number of PCs at which the curve begins to flatten out)

Methods: K-means Clustering

- Goal is to iteratively partition the data set into K pre-specified, non-overlapping clusters, where each data point belongs to only a single cluster
- Algorithm aims to make the intra-cluster data points as similar as possible but keep the clusters far apart for differentiation
- Every data point is assigned to the cluster with the nearest centroid by Euclidean distance
- Algorithm stops when all centroid values stop changing (data point assignments remain the same)
- Variant of the EM algorithm
  - K-means uses hard assignment based on convergence and uses the L2 norm for optimization
  - EM algorithm uses soft, probabilistic assignment and depends on the expectation of a point belonging to a particular cluster rather than on the L2 norm

Methods: Gaussian-Mixture Model

$$L(X_1, ...X_n; \mu_1, ...\mu_K; \Sigma_1, ...\Sigma_K; \pi_1, ...\pi_K) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

- Commonly used to investigate normally distributed subpopulations within a larger population
- Unsupervised technique, since data point assignment to each subpopulation is unknown
- Often fairly efficient with large data sets due to their similar computational and theoretical properties to standard Gaussian models
- Consists of multiple $k = 1, ..., K$ Gaussians, where K denotes the number of clusters
- Each Gaussian has a center (μ_k), a covariance (Σ_k), and a mixture probability (π_k), where $\sum_{k=1}^{K} \pi_k = 1$
- When the number of clusters in known, the EM algorithm can be used to estimate the GMM parameters

Methods: EM Algorithm

- Unsupervised method used to overcome the challenges of standard MLE procedures when dealing with mixture models with many parameter
- Can only be used if closed form expressions are available for updating the model parameters
- Has helpful property that the ML of the data continues to increase at each iteration (local maximum or saddle point is always guaranteed)
- Three steps:
  1. Initialization: K different random data points selected as cluster means and all cluster variances set to the overall sample variance
  2. E-step: Expectations of the cluster assignments for each data point are calculated, given all model parameters
  3. M-step: All expectations from E-step are maximized with respect to all model parameters, and all parameters are updated
- This process continues iteratively until convergence of all parameter estimates is achieved

Methods: (Divisive) Hierarchical Clustering

- Starts by dividing the entire data set into two homogeneous clusters and then similarly partitioning the data with each cluster iteratively until clusters produce identical items
- Does not require a pre-specified K value, but depends largely on the distance and linkage methods chosen for measuring inter- and intra-cluster distances
  - Spearman correlation distance and complete linkage were selected
- Data was log2-transformed and scaled up by 0.25 to address excess zero values

Methods: Silhouette Scoring

- Used to obtain the optimal K for clustering
- Tests the goodness of clustering for the target sample by measuring the average intra-cluster (a_i) and inter-cluster (b_i) distances for each observation

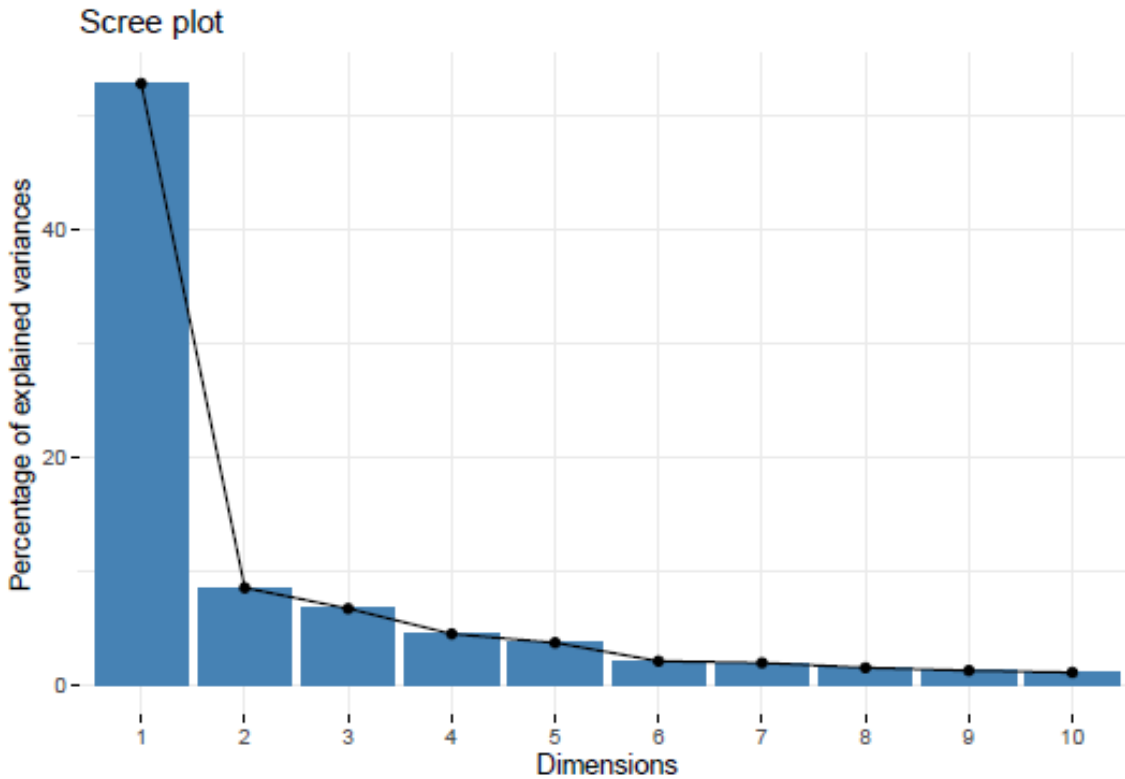
$$SilhouetteScore = \frac{b_i - a_i}{max(a_i, b_i)}$$

- Scores range from -1 to 1, where values closer to 1 indicate better clustering

Results: PCA

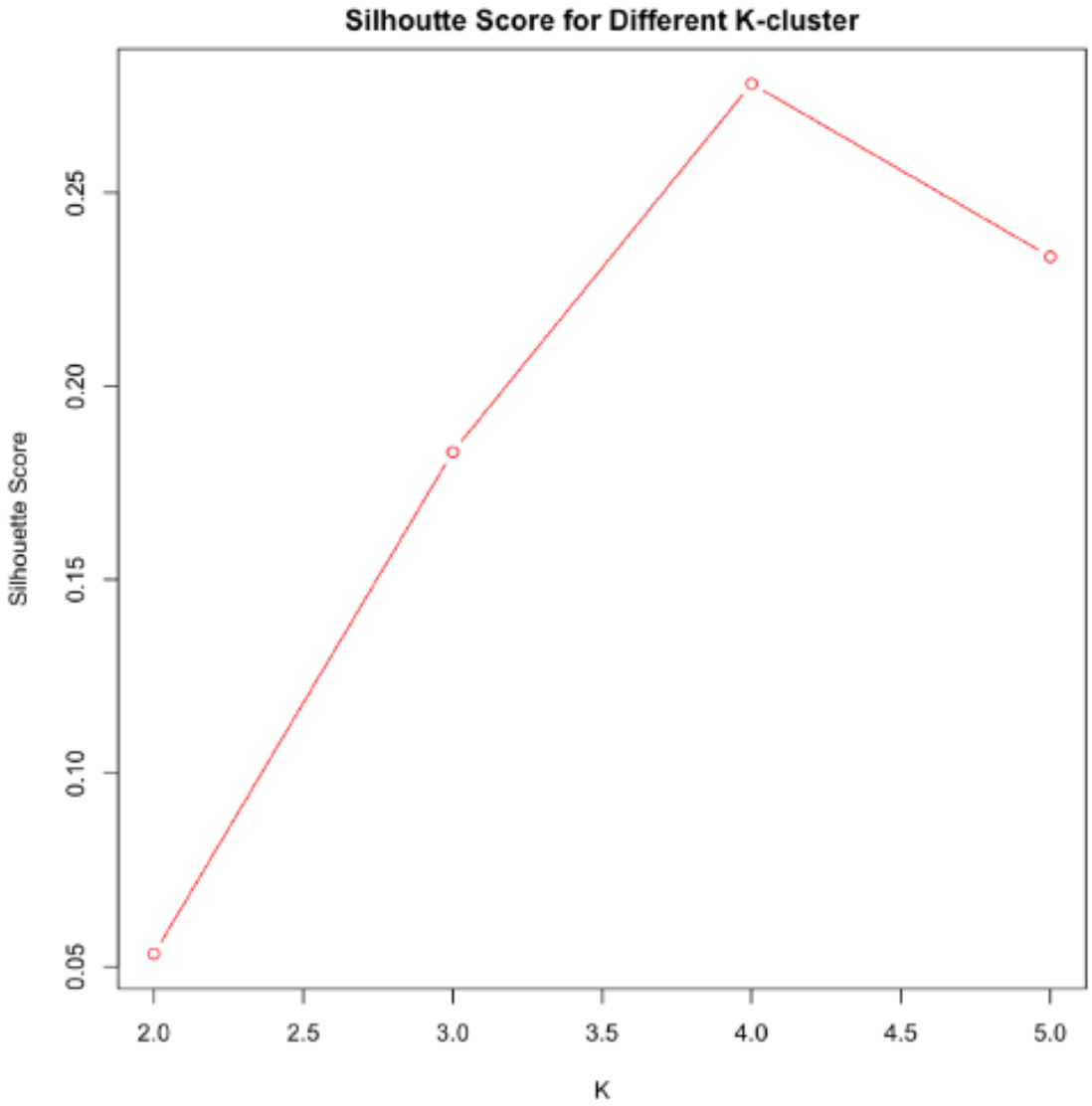
The 80% cumulative variance criterion, Kaiser's rule, and the Scree plot yielded 7, 34, and 6 PCs, respectively **(we proceed with 6 PCs)**

	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



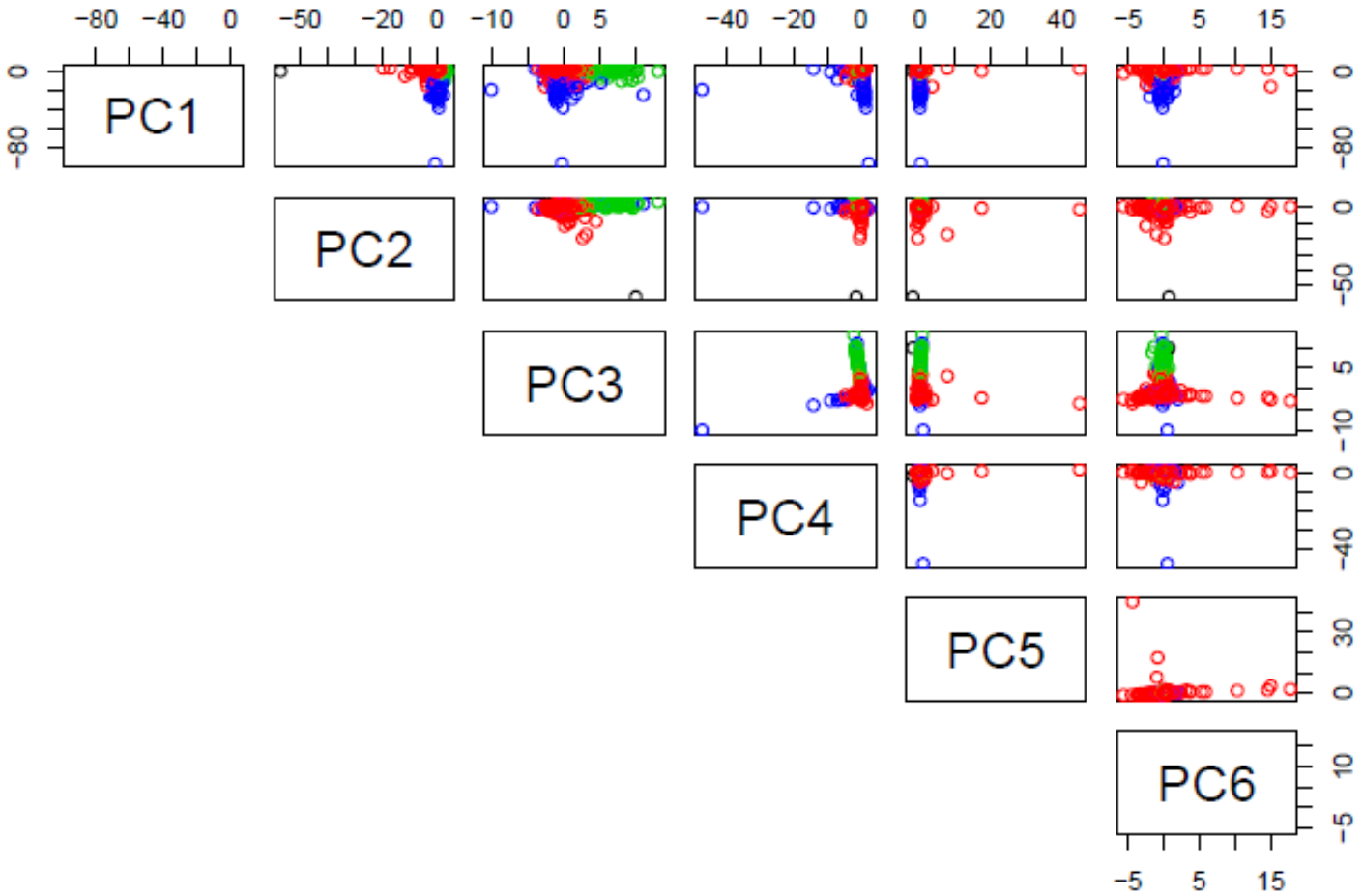
Results: Optimal Clusters

The largest silhouette score was obtained for K=4 clusters **(optimal)**



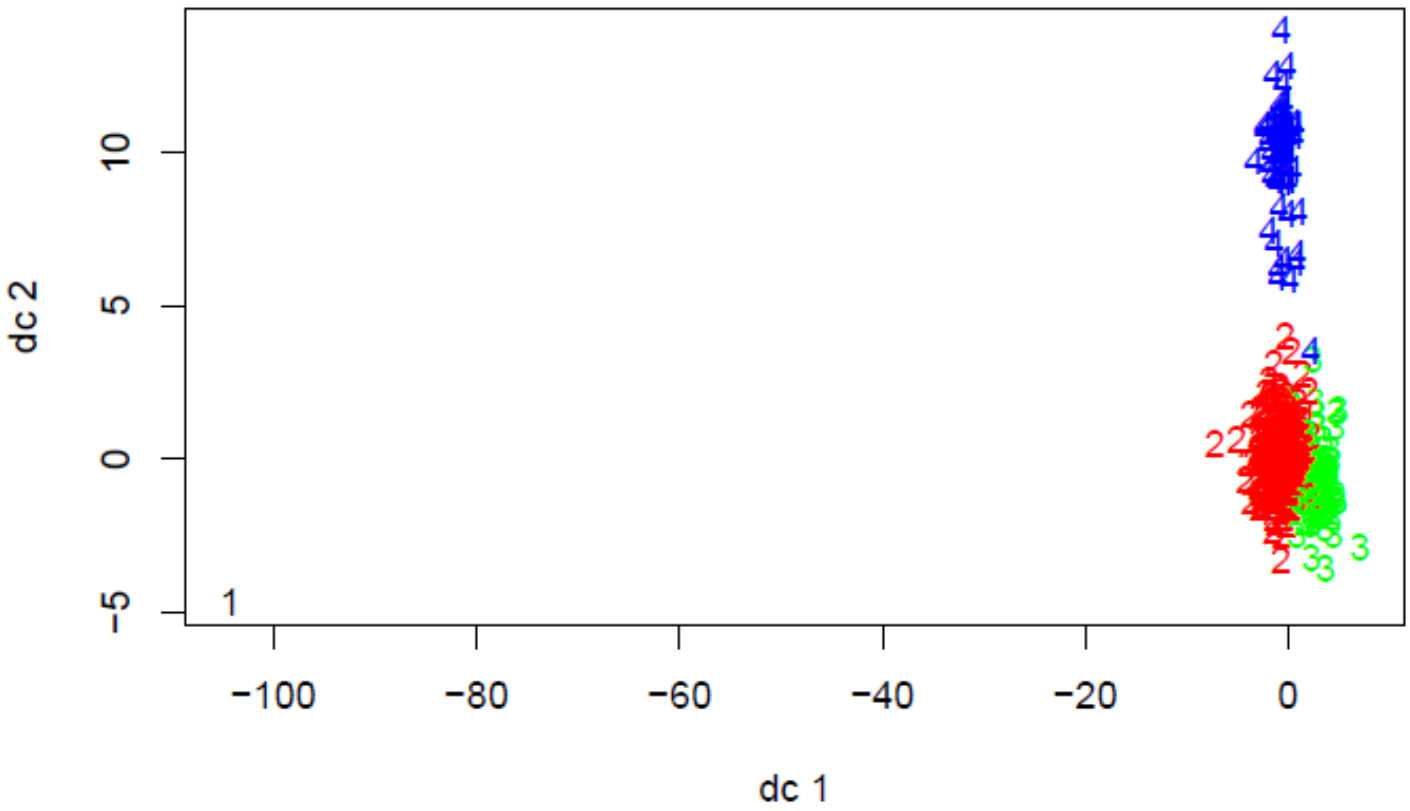
Results: Cell Assignments

Clustering results using 6 PCs, K=4 optimal clusters, and GMM with EM



Results: Cell Assignments

Most cells assigned to cluster 2 and one cell assigned to cluster 1



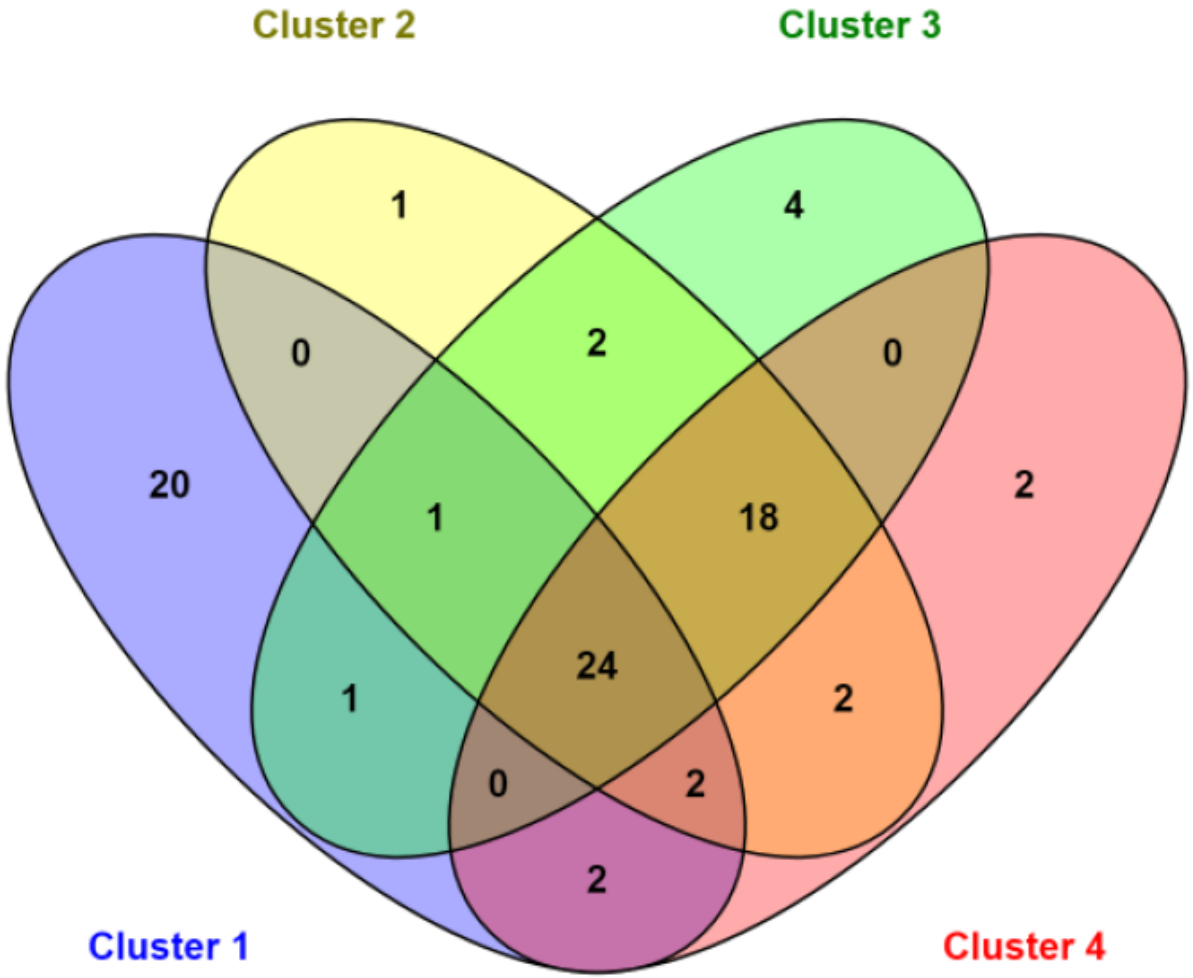
Results: Gene Expression Signatures

Among the top 10 genes most expressed across cells, there was some overlap between the clusters

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

Results: Gene Expression Signatures

There was more considerable overlap among the top 50 most expressed genes in each cluster across all cells



Results: Genes for Cluster Differentiation

We postulate that the most expressed genes that are unique to each cluster are the most important for differentiation

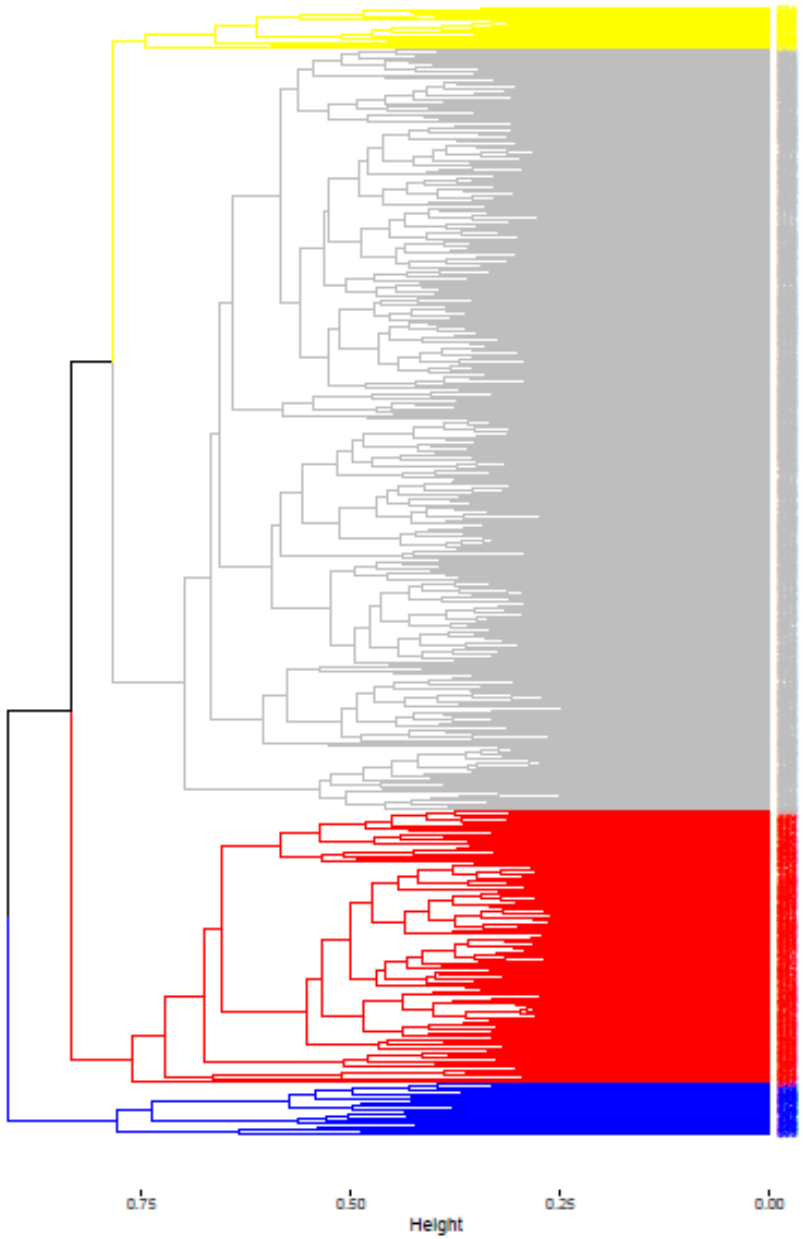
Cluster 1: Lum, Col8a1, Dcn, Mfap5, Mgst1, Ndufa4, Pi16, Tnc, Lox, Glipr1, Ccl2, Ccnb2, Prelp, Csn3, Sod3, Tspan7, Trf, Lrrc15, Sfrp2, Cpxm2

Cluster 2: Ly6e

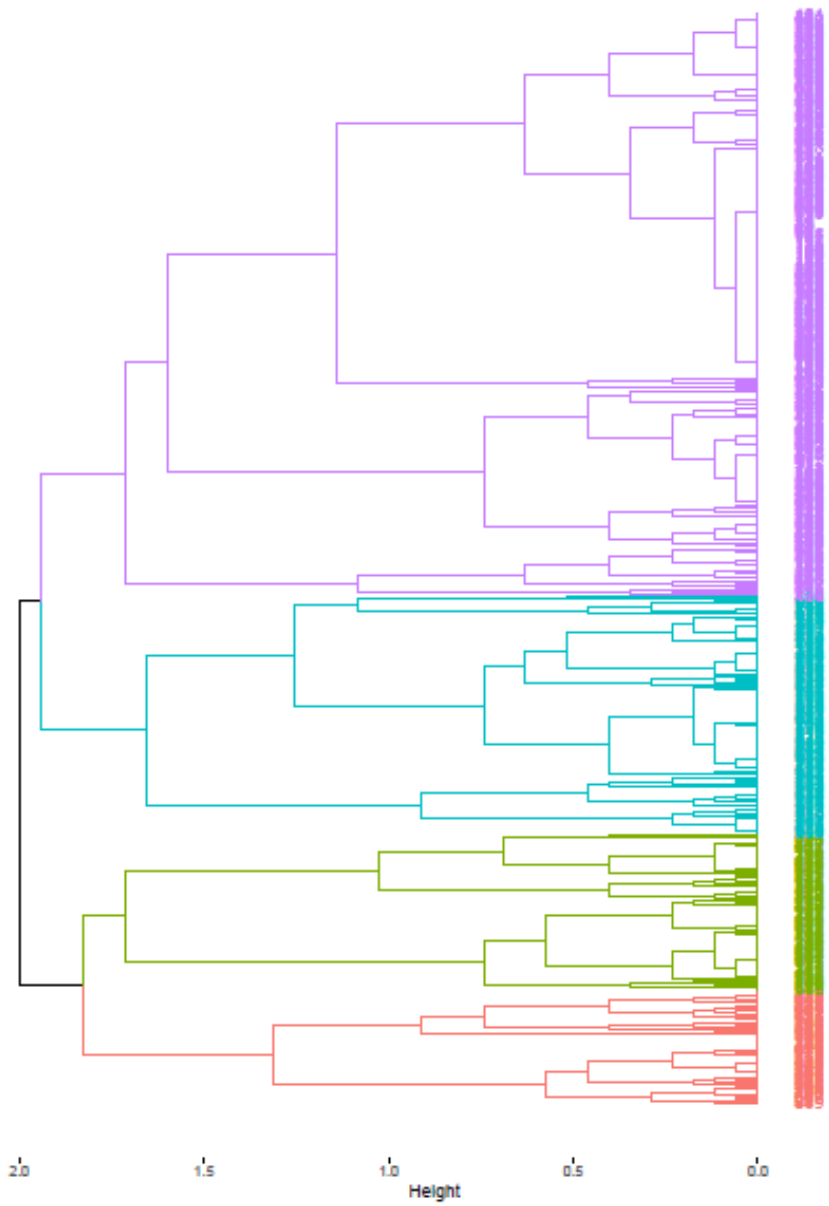
Cluster 3: Rgs16, Tinagl1, H2.M9, Apoa1bp

Cluster 4: Thbs2 and Nupr1

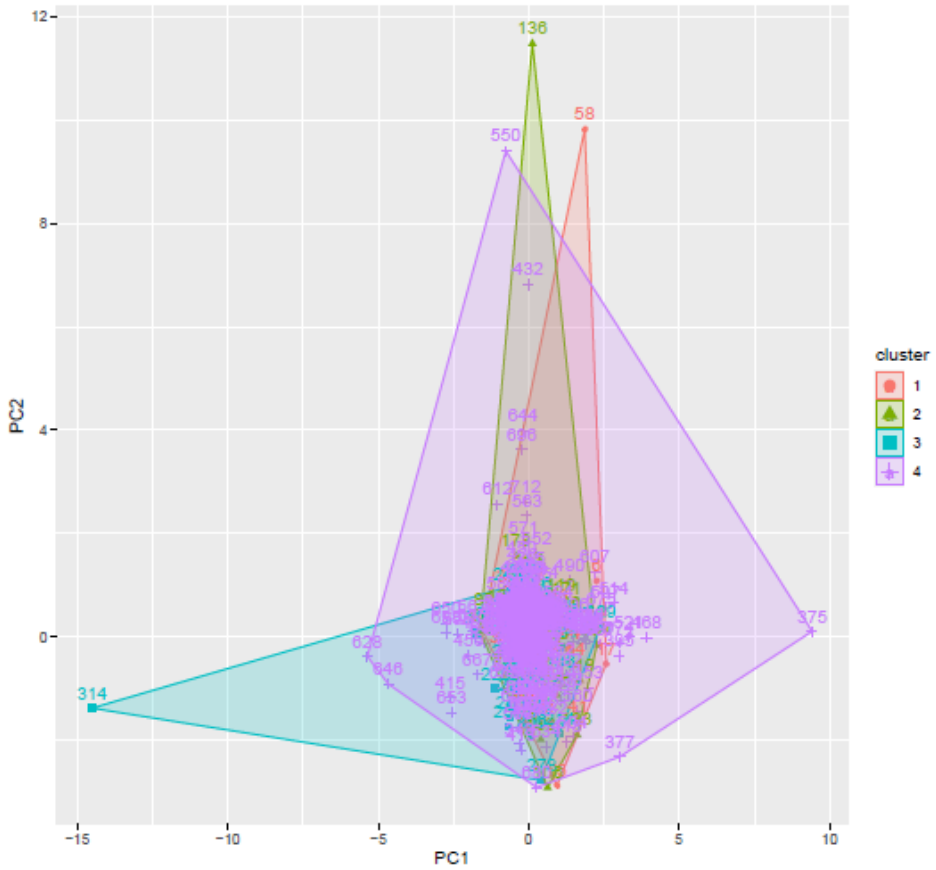
Results: Hierarchical Clustering

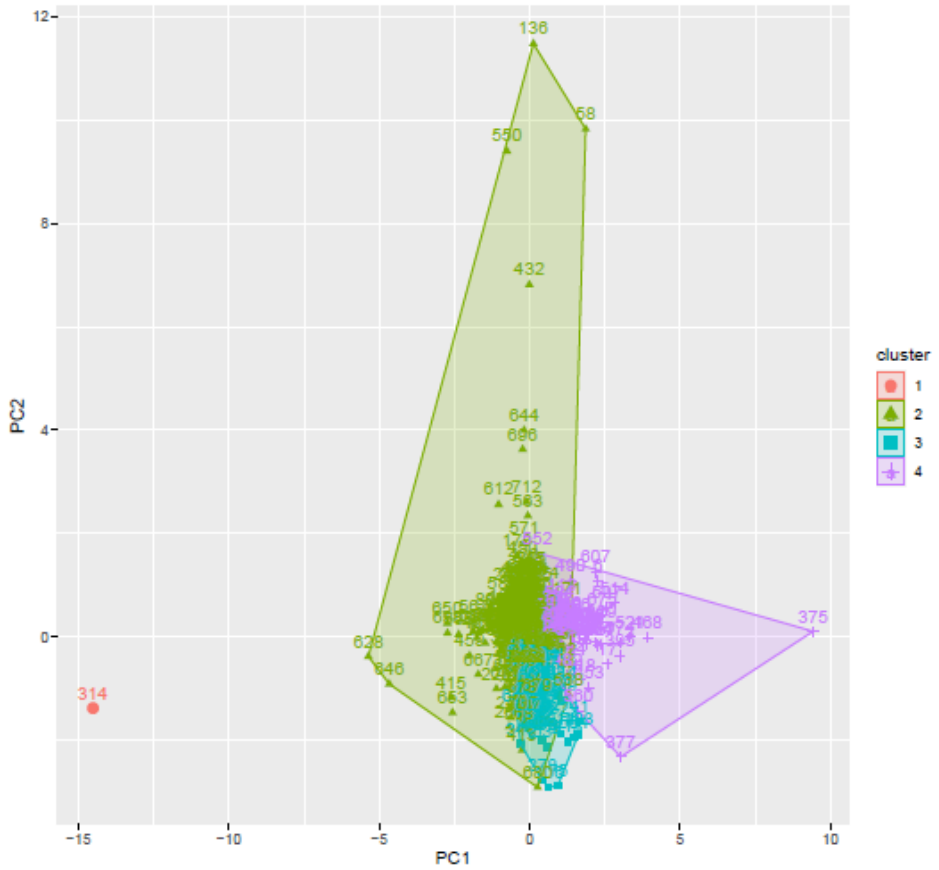


Results: Hierarchical Clustering with PCA



Results: Comparing Hierarchical and K-means Clustering





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

Discussion

- 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 address data with excess zeros

K-means:

- Shows higher efficiency in grouping homogeneous data, is more flexible in handling outliers, and is able to produce tighter clusters
- But initial starting value for K strongly impacts clustering results, making it less useful for high-dimensional data
- Overly condensed and unstructured plots of cluster assignments

Hierarchical Clustering:

- Does not require pre-specification of K, decreasing analysis bias
- Sensitive to outliers
- Dendrograms are very informative, showing dissimilarity metrics for each individual cell

*Both approaches are not suitable here, and hierarchical clustering performed even worse than K-means

End

Thank you!