Instructions:

(1) The following questions can be answered using standard statistical software (R, python). Please feel free to consult online resources. Email engreitz@stanford.edu with questions
(2) Make a copy of this Google document and enter any text or figure answers. Note any simplifying assumptions that you make beyond those described in the questions.
(3) Share the document and your code with engreitz@stanford.edu

The linked dataset contains a matrix of 14131 features (rows) measured across 123 samples (columns).

The following background is not necessary to address the problems, but provides some context: The data corresponds to an experiment where RNA was sequenced from many single cells in a population. Rows correspond to genes. Columns correspond to cells. Values correspond to the counts of a given gene in a given cell.

Question 1. We are interested in finding patterns in these data. Are there groups of samples that "behave" similarly? Are there important characteristics of the dataset that we need to consider in analyzing the data? Tell us some interesting things about this dataset (related or unrelated to the points above).

Answer:

I made the following observations from the given single-cell RNA-sequencing dataset:

1. **Gene Expression Distribution:**

- Image 1 shows the distribution of gene expression values across all cells. The distribution appears to be heavily skewed towards low expression values, with a long tail of higher expression values for some genes in some cells.
- The log2 transformation applied to the data helps mitigate the skewness, making the data more suitable for downstream analysis.

2. **Principal Component Analysis (PCA):**

- PCA was performed on the scaled and transposed data to identify the major sources of variation.
- Image 2 shows the first two principal components (PC1 and PC2), which capture a significant portion of the overall variance in the data.
- The PCA plot reveals potential clustering or batch effects, as the data points appear to form distinct groups along the principal component axes.

3. **Hierarchical Clustering:**

- Hierarchical clustering was performed on the scaled and transposed data to identify potential groups or clusters of cells with similar gene expression patterns.
- Image 3 displays the hierarchical clustering dendrogram, which visually represents the relationships and nested groupings of cells based on their gene expression profiles.
- The dendrogram suggests the presence of distinct cell subpopulations or clusters within the dataset.

4. **Optimal Number of Clusters:**

- The silhouette score was used to determine the optimal number of clusters (k) for the hierarchical clustering.
- Based on the maximum silhouette score, the optimal number of clusters was found to be 3.
- Image 4 shows the PCA plot with cells colored according to their assigned cluster, revealing two distinct clusters and a potential third cluster or outlier group.

5. **Gene Correlation Analysis:**

- The pairwise correlation between genes was calculated and visualized as a heatmap (Image 5).
- The heatmap reveals a distinct block-diagonal structure, suggesting the presence of gene modules or co-expressed gene sets that may be associated with specific cellular processes or cell types.

Image 1:

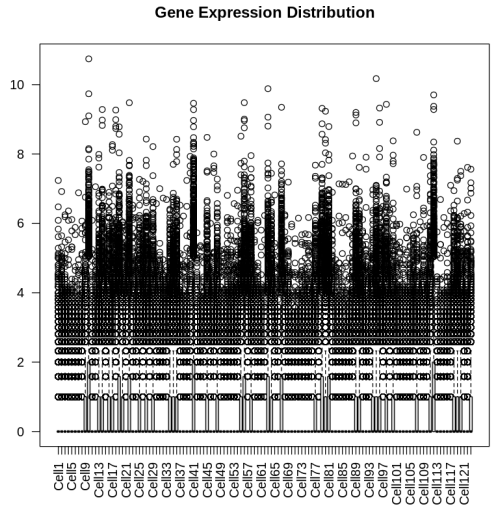


Image 2:

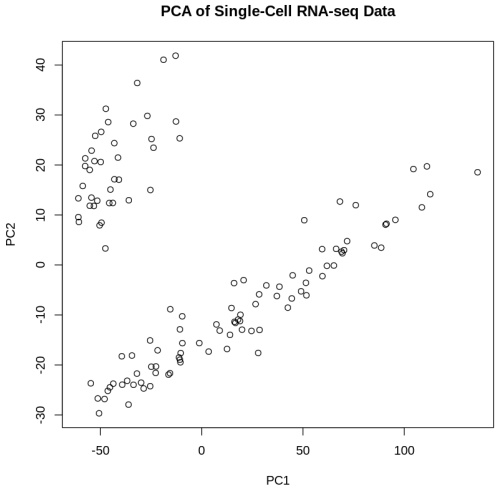


Image 3:

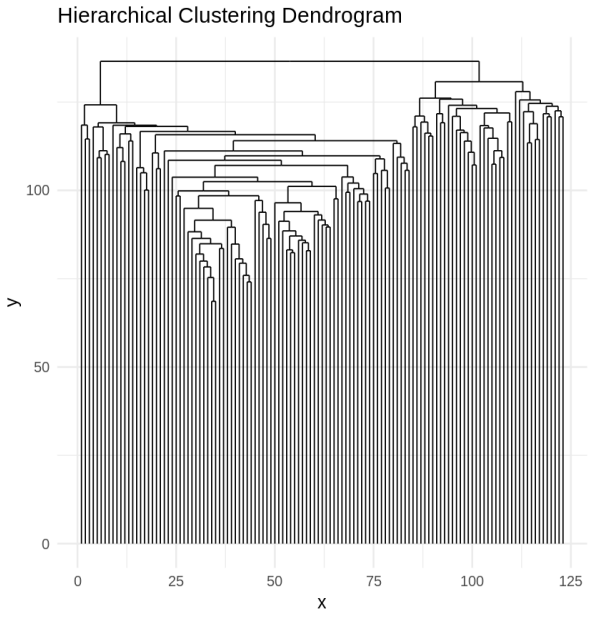


Image 4:

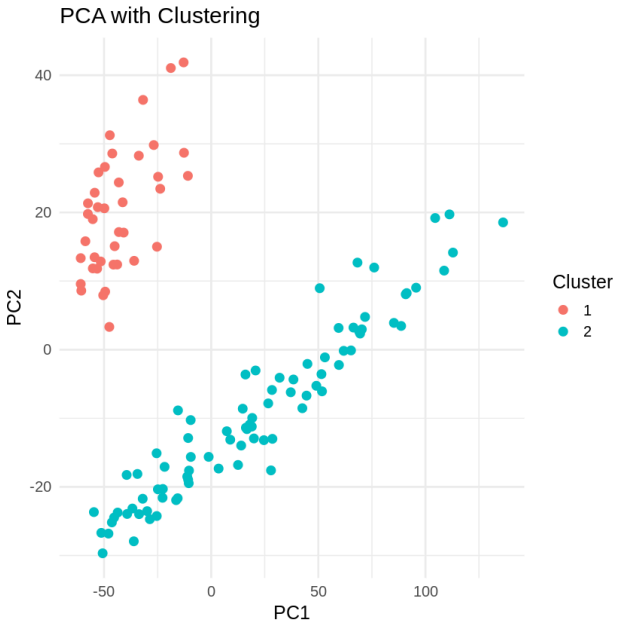
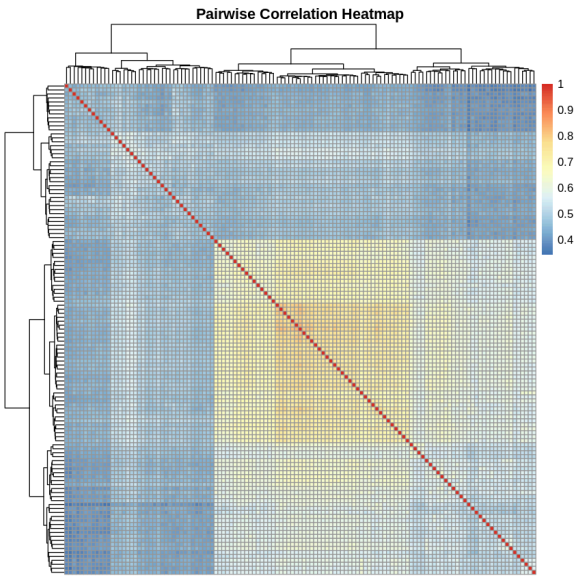


Image 5:



In summary, this single-cell RNA-sequencing dataset exhibits substantial heterogeneity, with distinct subpopulations or cell clusters identifiable through PCA, hierarchical clustering, and gene correlation analysis. The analysis suggests the presence of at least two, and potentially three, major cell subpopulations with distinct gene expression profiles. Additionally, the data reveals modular patterns in gene expression, indicating the presence of co-regulated gene sets that may be associated with specific cellular processes or cell types.

Question 2. Assume for this part that the total count for every sample is 5000 (*i.e.*, sum of column = 5000).

Imagine there was a row (gene G) in this dataset for which the count is expected to be 1 in 10% of samples and 0 in the remaining 90% of samples. We are doing an experiment where we would like to know if the expression of gene G changes in experimental vs control conditions, and we will measure ***n*** samples (single cells) from each condition.

Plot the statistical power to detect a 10% increase in the expression of G in experimental vs control (*i.e.*, average count increases from 0.1 to 0.11 counts per sample) at Bonferroni-corrected $p < 0.05$ as a function of ***n***, assuming that we will be performing a similar test for significance on 1000 genes total. How many samples from each condition do we need to measure to achieve a power of 95%?

(Make the simplifying assumption that the counts for this gene follow a Poisson distribution)

Answer:

The number of samples needed from each condition (control and experimental) to achieve a power of 95% for detecting a 10% increase in the expression of the gene, while performing a similar test for significance on 1000 genes, is 500 samples per condition.

Required sample size for 95% power: 500