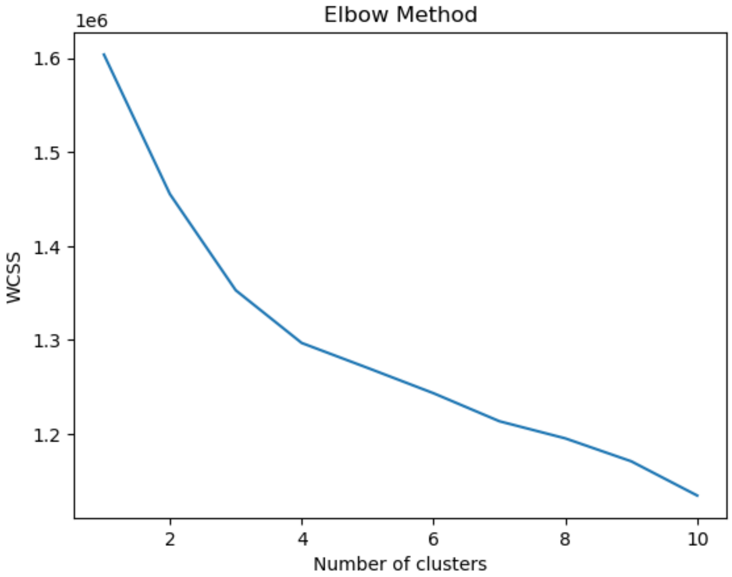


Figure 1: Elbow plot suggests that the optimal cluster number is n=4. The optimal point occurs when Within Cluster Sum of Squares (WCSS), or the sum of the squared distance between each member of the cluster and its centroid, begins to level off.

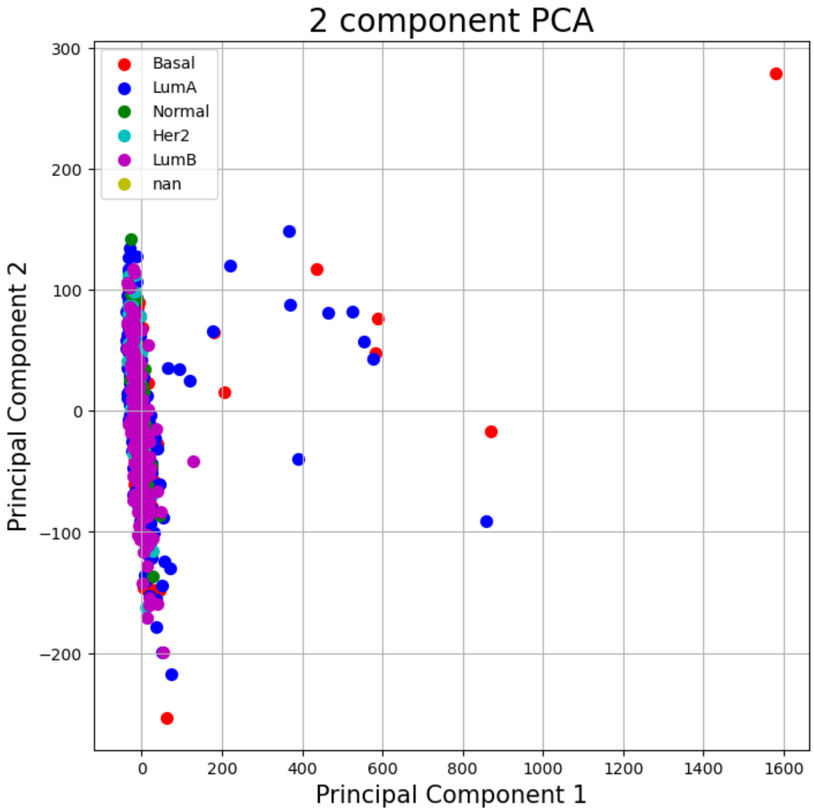


Figure 2: PCA dimensionality reduction fails to separate patients in the RNA dataset based on their BRCA subtype. Most subtypes are virtually indistinguishable from each other, while a few individuals of Basal and Luminal A subtypes show some deviation from the majority.

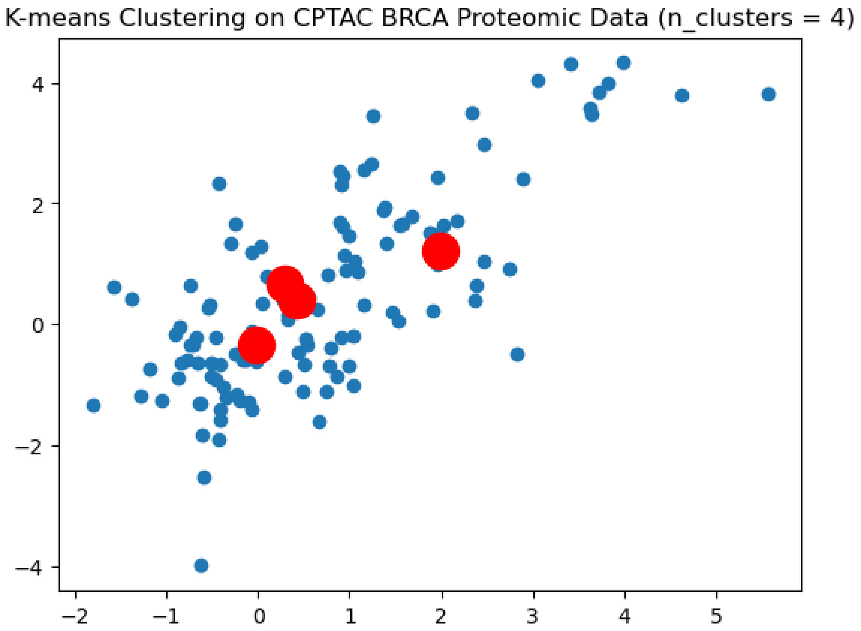


Figure 3: K-means clustering applied to CPTAC BRCA proteomic data identifies 4 cluster centroids (red). Dispersion of data points in relation to clusters is non-uniform, and overlap between centroids is seemingly high.

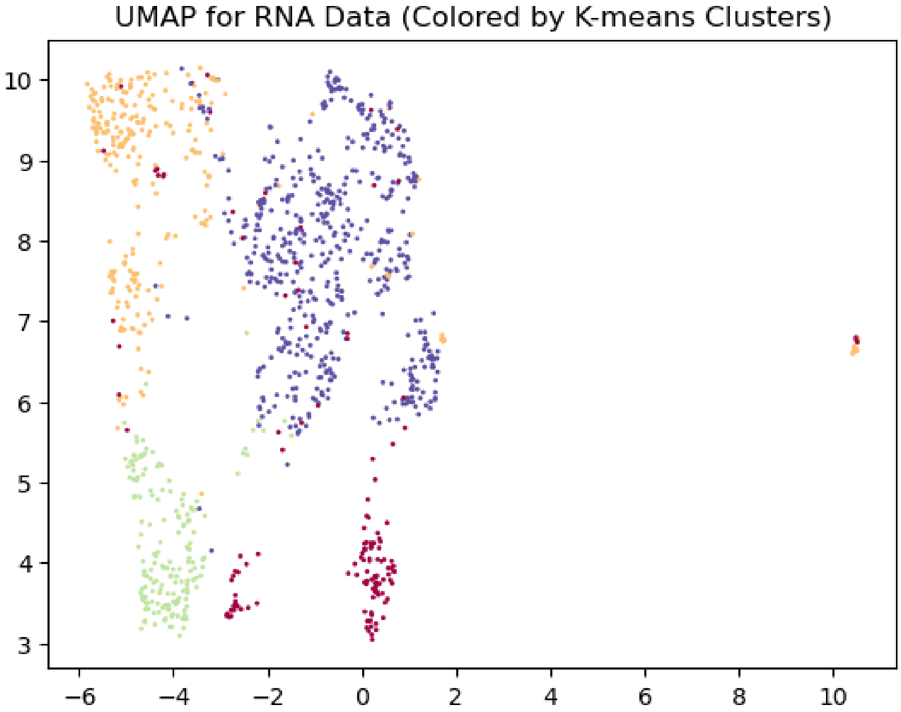


Figure 4: The four K-means clusters effectively segregate the patients in the RNA dataset into four distinct groups. The green subgroup is the most distinct and concentrated, while the yellow, purple, and red subgroups are more dispersed and exhibit increased overlap with other clusters.

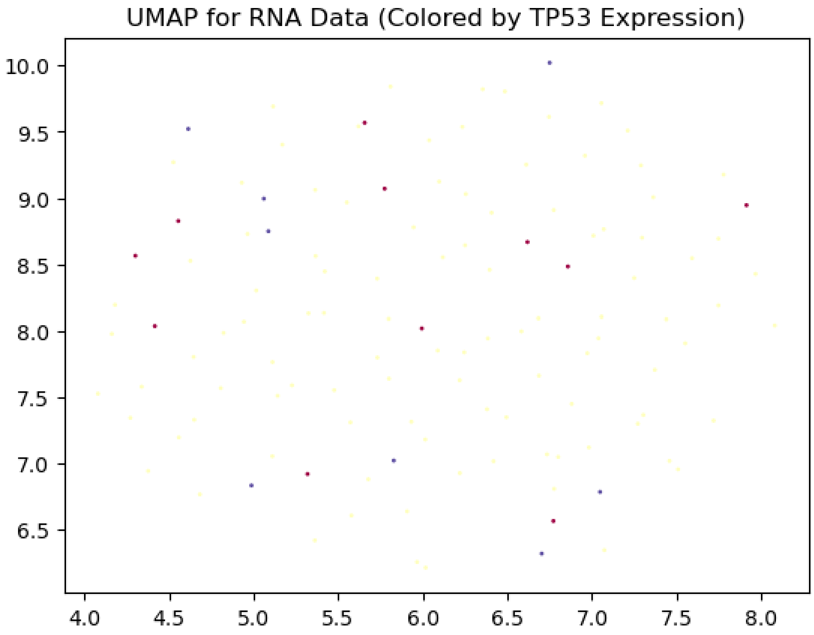


Figure 5: The expression of the TP53 gene (Over, Average, Under) is not effective at distinguishing between different groups of patients in the RNA dataset. Possible errors in visualization are indicated by the sparsity of data points.

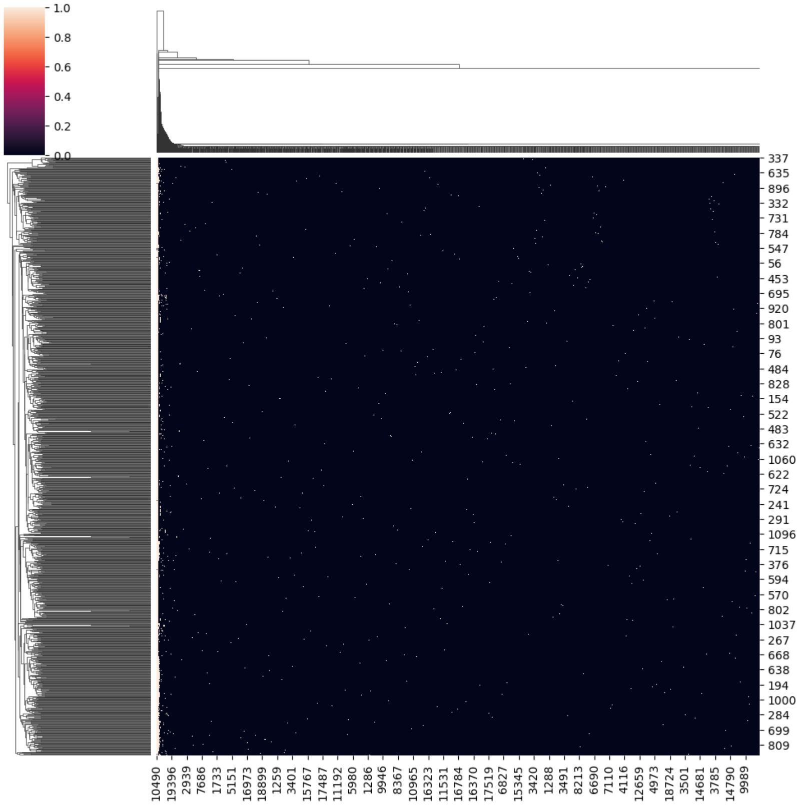


Figure 6: Hierarchical clustering heatmap using RNA data signifies little similarity in data points between and within clusters.

Questions:

1. Looking over your figures, does anything surprise you? Why or why not?

   Referencing Figure 2, I'm surprised that PCA dimensionality reduction techniques are unable to stratify the patients in the RNA dataset. I expected that the patients in the RNA clinical dataset would be easily distinguishable based on demographic and cancer-specific variables, but this turned out not to be the case.

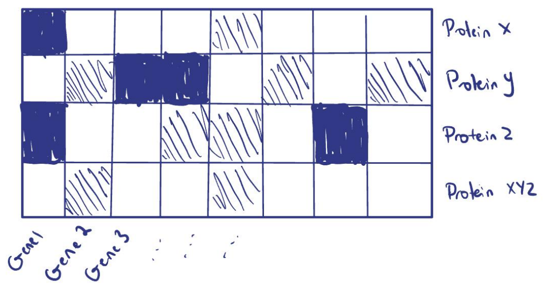
   Referencing Figure 3, it's interesting that the K-means algorithm identified centroids that were so close to one another as opposed to ones that are more uniformly distributed. I had originally expected the centroids to be more uniformly distributed as the K-means algorithm aims to minimize within-cluster Euclidean distances. The results seen in the figure may simply be a result of the inherent distribution of the dataset.

2. Now that you have clusters, what information would you like to know about each cluster? How would you get this information?

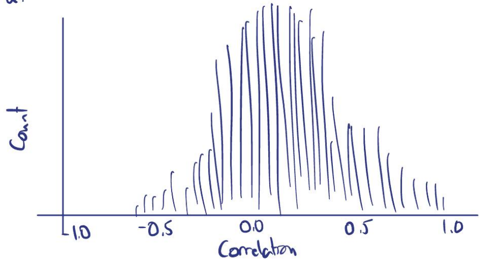
   With clusters of distinct data points, for example those depicted in Figure 4, it would be interesting to explore the difference in data point distribution based on clustering criteria. For example, are there certain overlaps between the groups formed when clustering by cancer subtype versus pathologic stage? In such an example the data would come from the clinical dataset, but data from other sources (ie. transcriptomic, proteomic, etc.) would also potentially be valuable here.

3. Brainstorm two ways you could combine RNA and Protein information into one figure. Provide two sketches of these figures.

1. RNA-Seq / Protein Expression Heatmap



2. Gene-wise correlation of RNA and protein expression


Count
Correlation