Liz Garcia Ovalles, Ekene Afulukwe, Victoria Li

Experiment Replication and Expansion – Clustering Cancer Gene Expression Data: A Comparative Study

Abstract

Clustering algorithms have been used to help discover cancer subtypes, but comparison of the different algorithms had not been performed before the replicated paper [1]. This study provides a comprehensive comparison of the algorithms to guide future algorithm selection for cancer subtypes research. The study evaluates and compares the clustering of cancer gene expression data using seven clustering algorithms and eight different proximity measures. The corrected Rand index (cR) assessed clustering performance. The replicated analysis is different from the original study, with k-means outperforming other methods and the finite mixture of Gaussians ranking second for Affymetrix data sets. For cDNA, spectral and shared nearest neighbors performed best. Furthermore, Manhattan distance yielded the best mean cR indices for Affymetrix datasets. In addition, analysis with PCA reduced performance, likely due to information loss. Tissue type and microarray technology also influenced clustering results, with blood tissue datasets achieving better classification and higher cR indices compared to brain tissue datasets and cDNA datasets displaying better classification than Affymetrix datasets in general. Better classification performance was observed for k-means clustering and PCA compared to hierarchical clustering on a selected blood tissue dataset (cDNA and Affymetrix). This study performs a recapitulation and expands on the original study by examining the impact on classification performance of an additional proximity measure (Manhattan distance), exploring datasets from different microarray platforms, and analyzing datasets containing diverse tissue types.

Background

The original paper [1] provides a comparative study of seven clustering algorithms along with the seven proximity measures applied to 35 cancer gene expression datasets. The seven clustering algorithms used are: single linkage, complete linkage, average linkage, k-means, finite mixture Gaussians (FMG), spectral clustering (SPC), and shared nearest neighbor based (SNN) clustering. The seven proximity measures are: Pearson's Correlation coefficient, Cosine, Spearman Correlation coefficient, original Euclidean Distance (Z0), standardized Euclidean Distance (Z1), scaled Euclidean Distance (Z2), and ranked Euclidean Distance (Z3). The goal of the study is to evaluate and compare the performance of the clustering methods in recovering the true structure of the data, using the corrected Rand index as a

measure of success. The study aims to identify general trends and guidelines for clustering cancer gene expression data, especially for future research in the field.

This work is influential because it provides a comprehensive, large-scale comparison of clustering algorithms applied to cancer gene expression data, offering valuable insights into the effectiveness of different methods. By establishing benchmarks and identifying general trends, it guides future research and algorithm selection in this field. Additionally, the proposed central repository [2] for clustering evaluations facilitates ongoing progress and comparison of methods, making it a key resource for bioinformatics.

In this project, we replicated all the provided figures in the original study as well as created additional figures for further investigation. Our goal was to assess the replicability of the results reported and expand on the previous work by including figures and analysis of our own with the available datasets.

Results

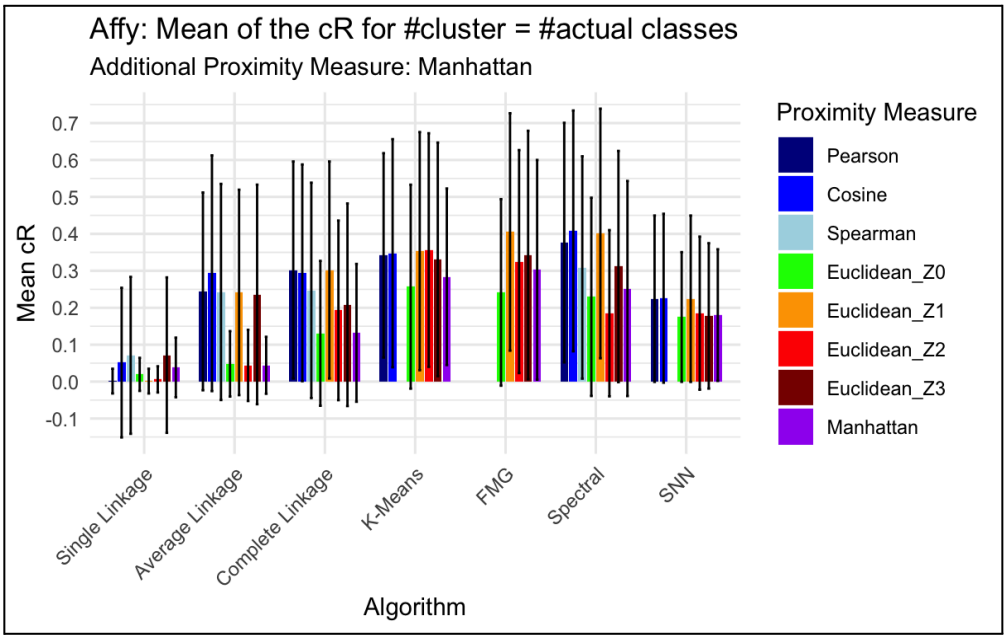
Victoria's Contribution:

Figure 1a: The published Figure 1a compares the mean cR index when k is the actual number of classes for the Affymetrix datasets. It showed that the k-means and FMG have the best performance, and the Euclidean measures had the lowest mean cR index across most algorithms. In my replicated Figure 1a, the results were similar. The hierarchical clustering algorithms reflect the published figure's, with slight differences in Euclidean distance. In Figure 1a, the Euclidean distance of Z1 performed better than other Euclidean distance transformations. Unlike the published figure, k-means performed slightly better than FMG. Another difference is that the performance of the Euclidean distance in SPC performed better. In the published figure, the mean cR index for Euclidean distances were low compared to Pearson, Cosine or Spearman, but my figure shows that Euclidean distance almost performed as well, such as the Z1 Euclidean distance. Lastly, the Manhattan distance performed similarly to the Euclidean distance.

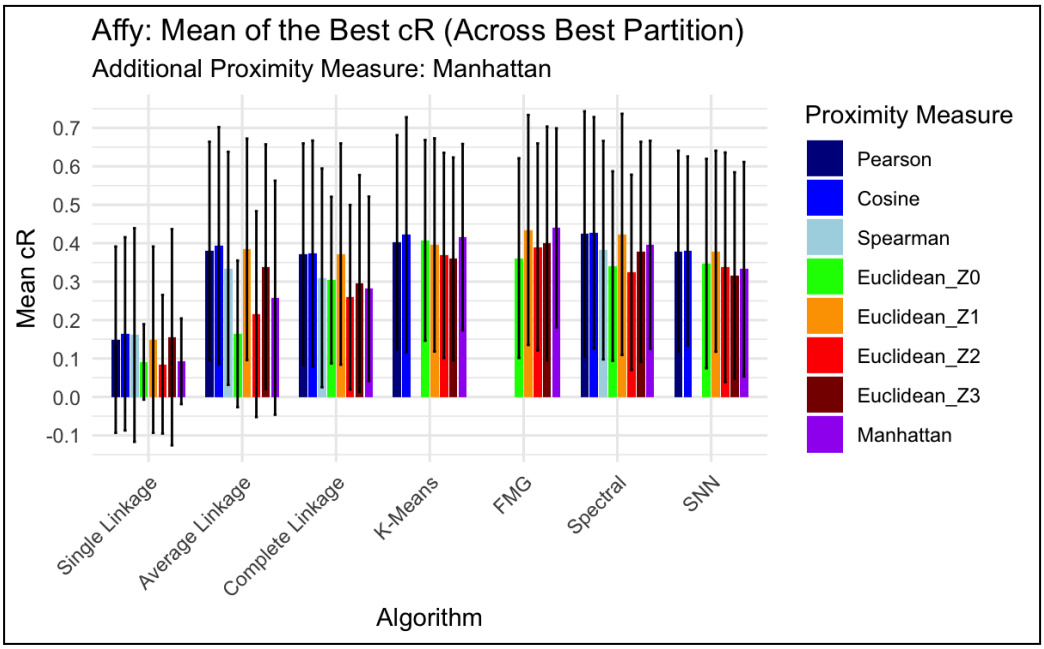
Figure 1b: The published Figure 1b compares the mean cR index for the best partitions for the Affymetrix datasets. It shows that the mean cR index improved from Figure 1a, which was reflected in my replication. Likewise, Manhattan distance improved in performance across all algorithms. In the published figure, it was more obvious that FMG performed the best, followed by k-means. However, my replicated figure shows that k-means, FMG, and SPC performed around the same. In addition, average linkage and complete linkage performed better than expected, close to k-means. My mean cR index in both Figure 1a and Figure 1b was still lower compared to the published Figure 1.

Figure 1

a)



b)

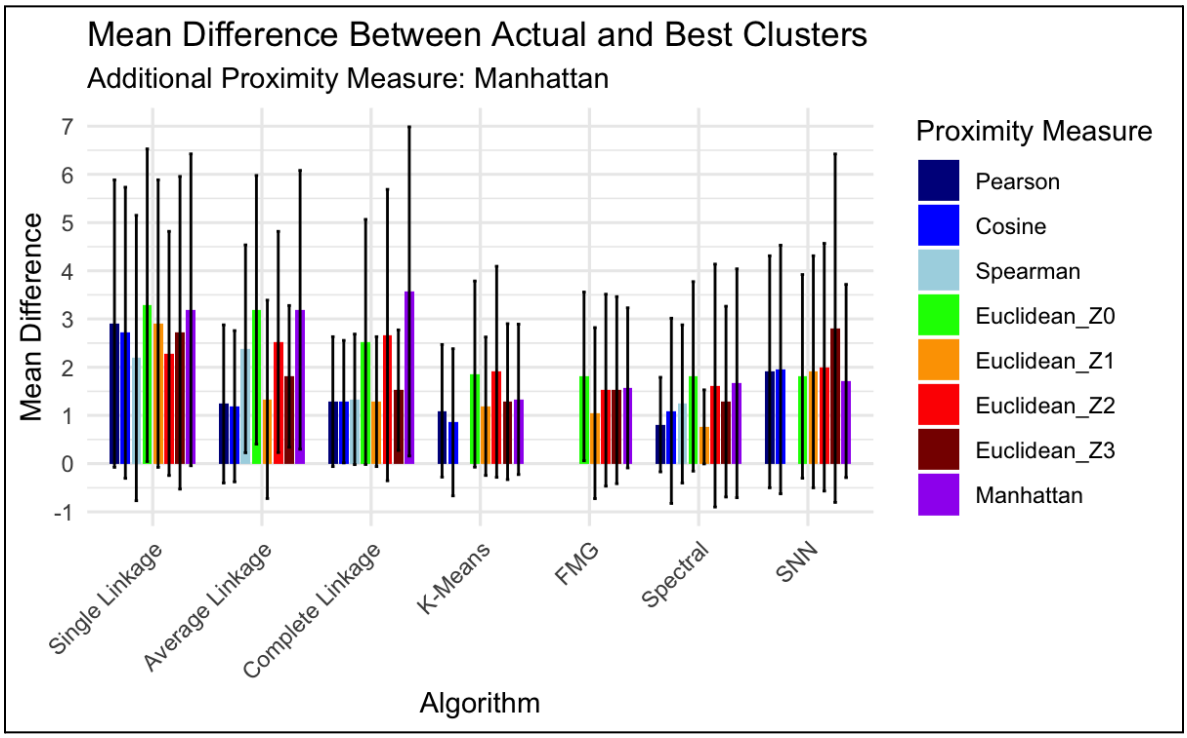


Affy: Mean of cR: *The mean of the cR for k equals number of actual classes (a) and mean of cR for best partition (b) is displayed. The missing bars indicate that the proximity measures were not compatible with the algorithm.*

— Victoria

Figure 3: In the paper, Figure 3 compares the number of actual classes versus the number of clusters produced by the algorithms for the Affymetrix dataset. The published figure shows that single linkage has the highest difference, reaching a difference of 4, while k-means, FMG, and SPC had the lowest difference of around 2. My replicated figure shows a similar result as the published figure, where the max difference is almost 4. In addition, k-means, FMG, and SPC also have the lowest differences. Lastly, we can also observe that across the hierarchical clustering algorithm, the Manhattan distance has the highest difference compared to other proximity measures used.

Figure 3

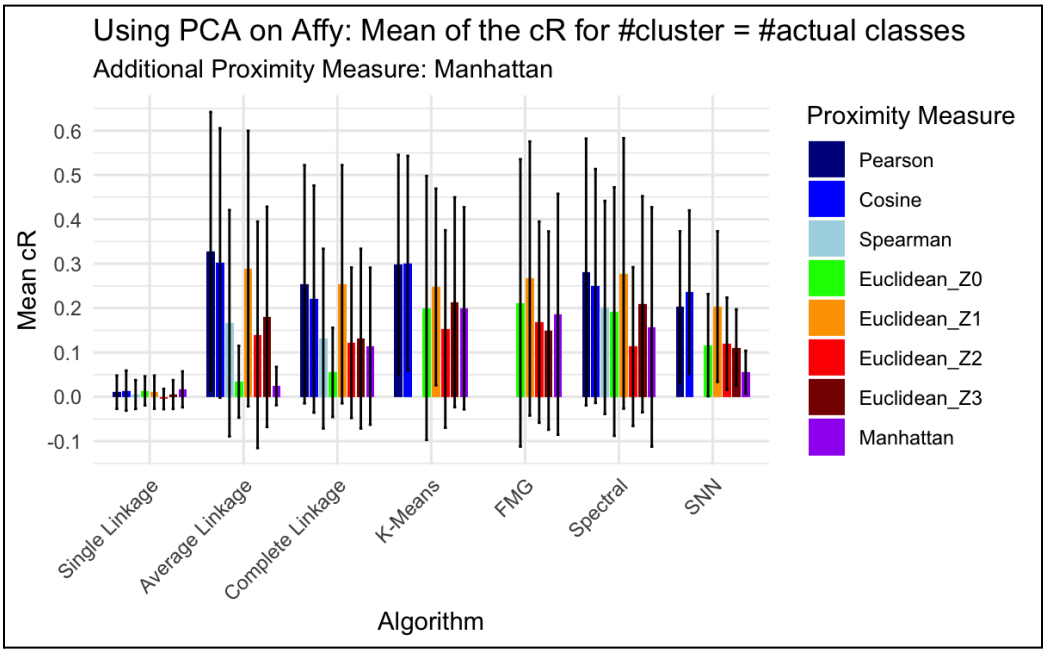


Mean Difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number of clusters from the best partition for Affymetrix datasets.* — Victoria

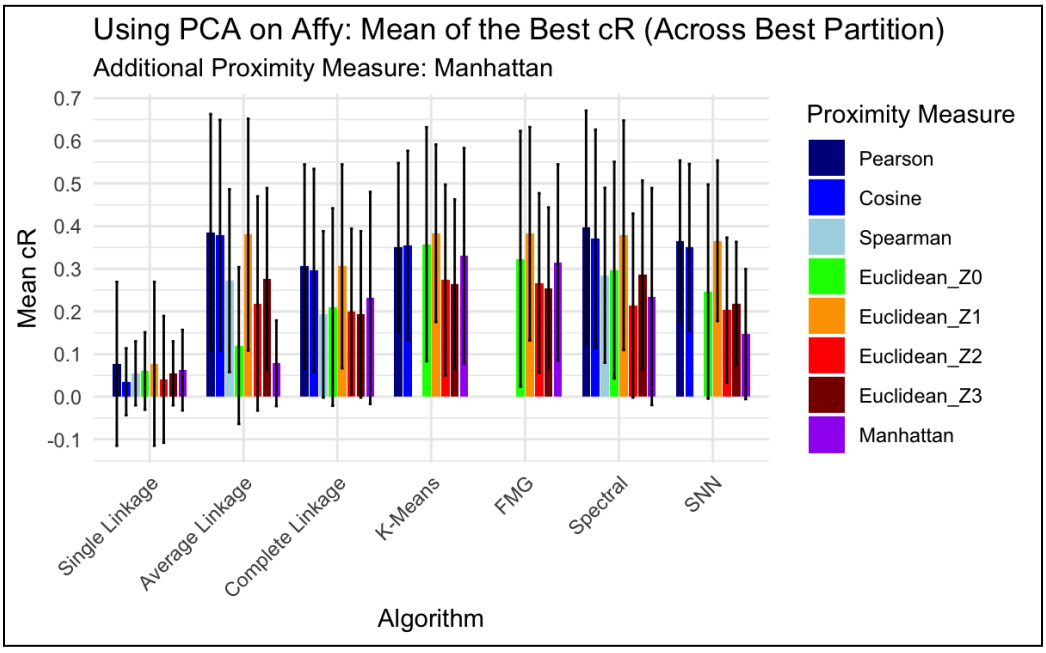
Figure 18: In this figure, I replicated the published figure 1a and 1b but with PCA on the Affymetrix datasets before applying the clustering algorithms and proximity measures. Overall, the general trend is the same as Figure 1 is retained where k-means, FMG, and SPC had the best performance. In addition, using the best partition also increased the mean cR index from Figure 18a to Figure 18b. However, the mean cR index decreased in comparison to Figure 1. The performance of average linkage improved with PCA, especially for the proximity measures of Pearson, Cosine, and standard Z1 Euclidean distance.

Figure 18

a)



b)

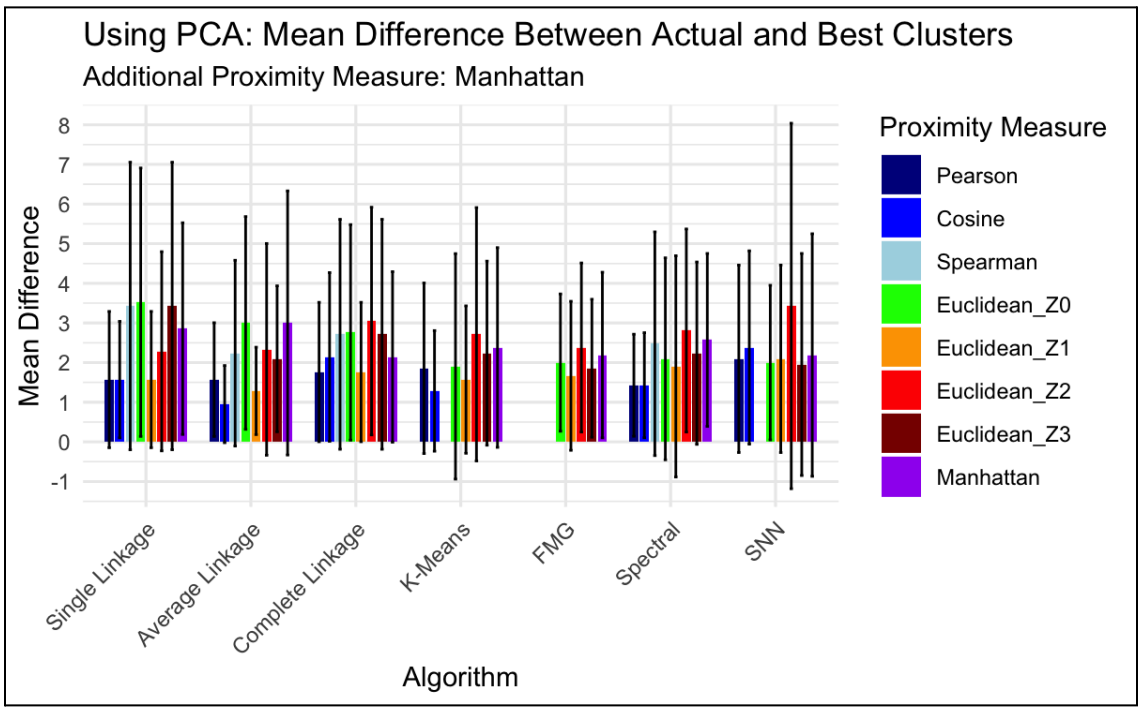


PCA on Affy:Mean of cR: *This displays the mean of cR when k equals actual number of classes (a) and mean of cR for best partitions (b) after PCA was applied to the Affymetrix datasets.* — Victoria

Figure 19: In this figure, I replicated the published figure 3 but with PCA. Likewise, the general trend still remains the same, where the hierarchical clustering had the highest differences, while K-means, FMG, and SPC had the lowest difference. The max difference was higher than the max difference found

in Figure 3. Generally, the Euclidean proximity measure and Manhattan distance had the highest difference across all clustering algorithms.

Figure 19

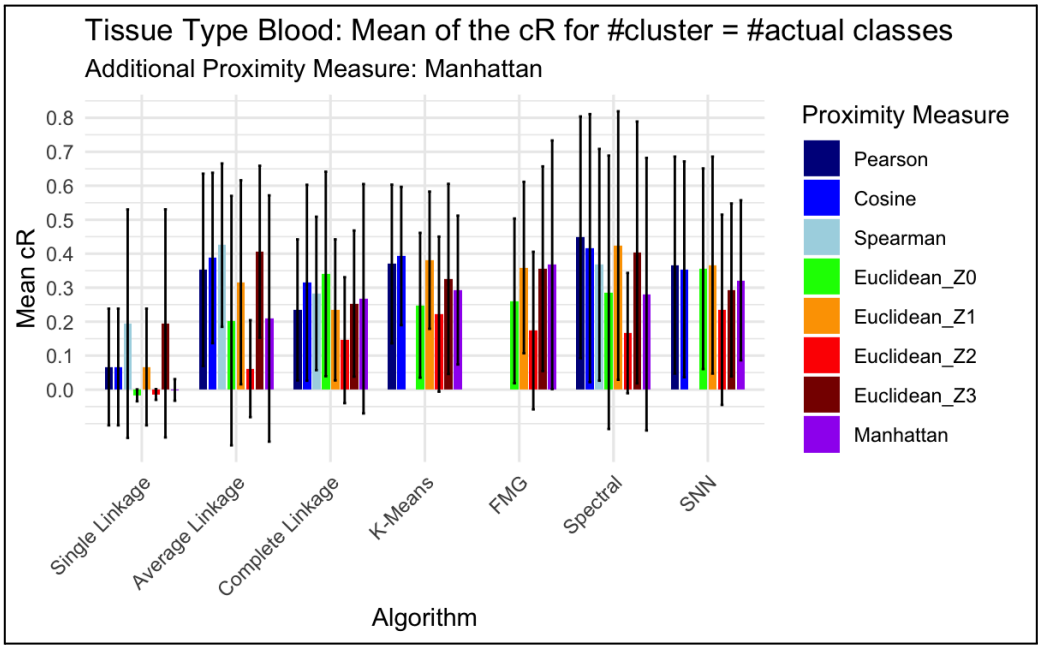


PCA: Mean difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number clusters from best partition after PCA was applied on the Affymetrix datasets.* — Victoria

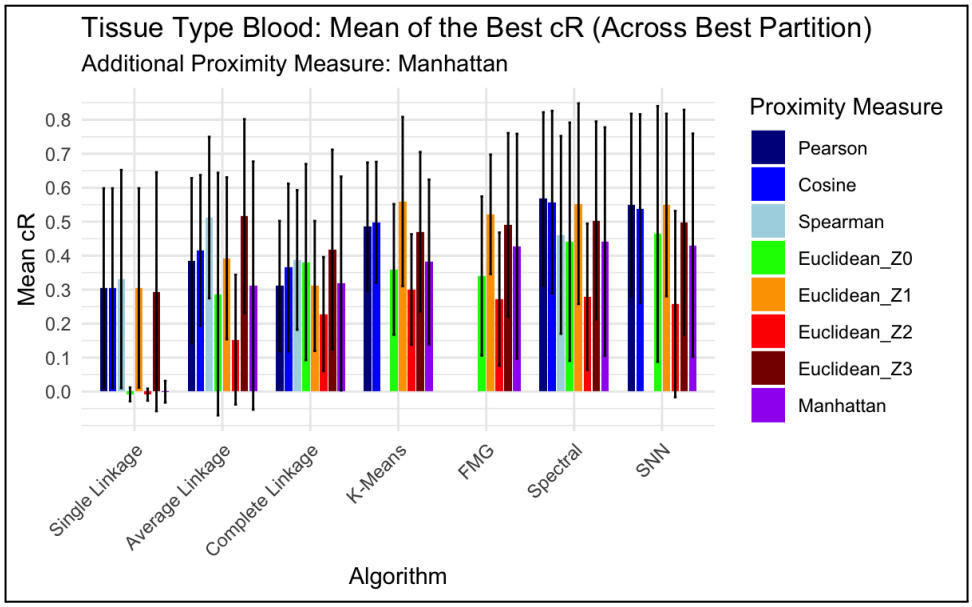
Figure 20: I replicated Figure 1 but with a focus on blood tissue datasets from both Affymetrix and cDNA datasets. The resulting Figure 20a shows that the mean cR index is lower compared to Figure 1a. The single linkage performed a lot worse, but average and complete linkage performed better than the original replication across the algorithms. However, in Figure 20b, the mean cR index was higher than Figure 1b. Specifically, performance of SPC and SNN clustering improved for certain proximity measures, like Pearson and Cosine. The standardized Z1 Euclidean distance also performed well and, in some cases, the best out of the proximity measures.

Figure 20

(a)



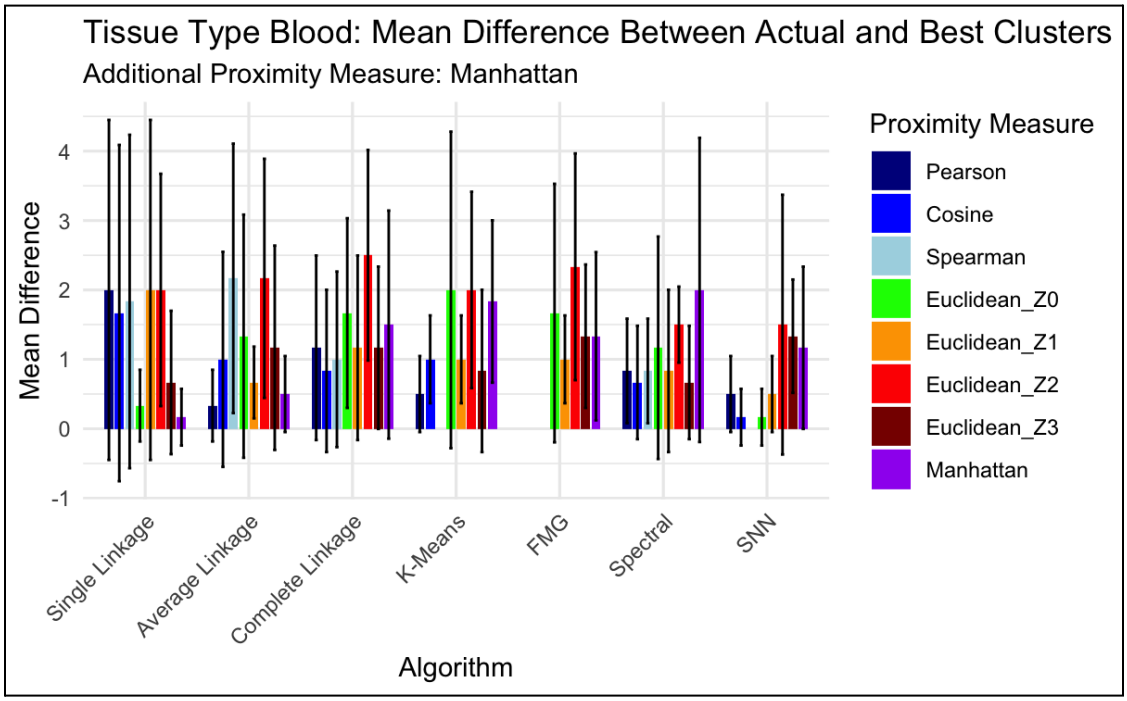
(b)



Blood Tissue: Mean of cR: *This displays the mean of cR when k equals actual number of classes (a) and mean of cR for best partitions (b) of the blood tissue datasets.* — Victoria

Figure 21: In Figure 21, I compared the mean of the differences in the actual number of classes versus best partitions again using Figure 20. The range of the difference is much smaller than the original replication of Figure 3, where the max mean difference is 2. In this figure, the scaled Z2 Euclidean distance had the most number of differences in almost all of the clustering algorithms.

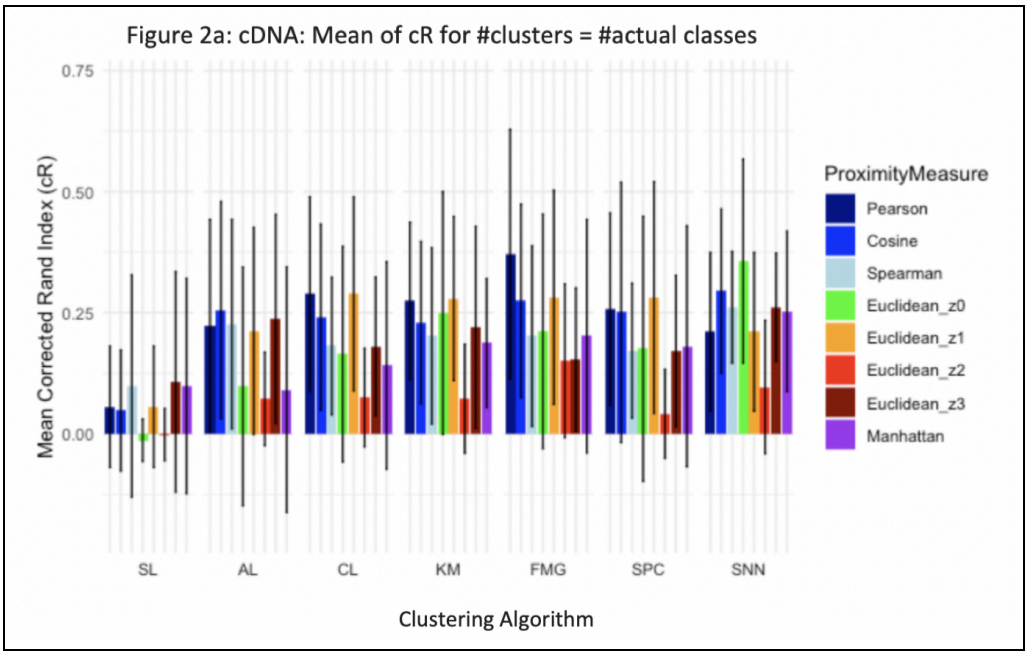
Figure 21

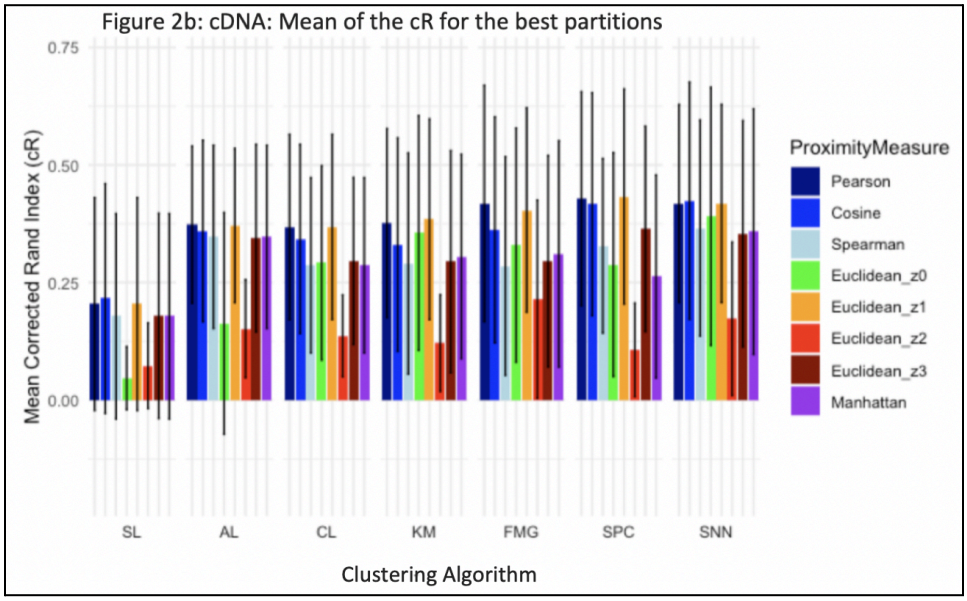


Blood Tissue: Mean difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number of clusters from the best partition of the blood tissue datasets.* — Victoria

Ekene's Contribution:

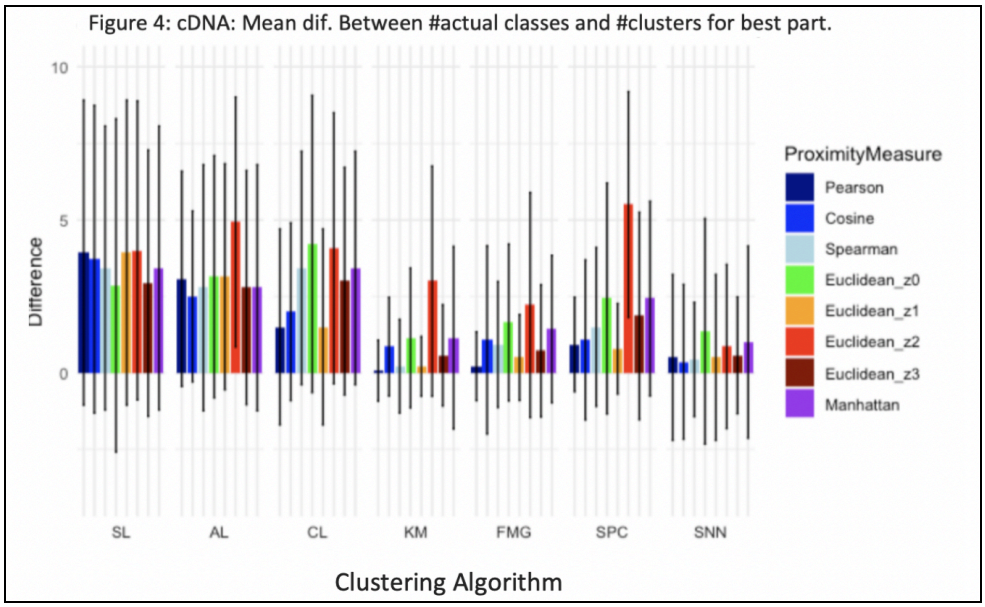
Figure 2a and 2b: In the published Figure 2a, K-means and FMG produce the greatest cR values, and the hierarchical clustering algorithms produce the lowest cR values. In my replicated Figure 2a, the cR values of all algorithms are more equal, except for single linkage which has lower values. FMG and SNN produce the greatest cR values in my figure. Additionally, in the published figure Pearson's correlation metric produces some of the greatest cR values for each clustering method and Euclidean_z2 produces some of the lowest, but in the replicated figure both distance metrics produce the highest cR values. The Manhattan distance, our added distance metric, never performs the best or worst for each clustering method. The cR values for the replicated figure are generally lower than that of the published figure. Both the published Figure 2b and the replicated Figure 2b show the same general trends that were observed in Figure 2a, but this an overall increase in cR values across all clustering methods.





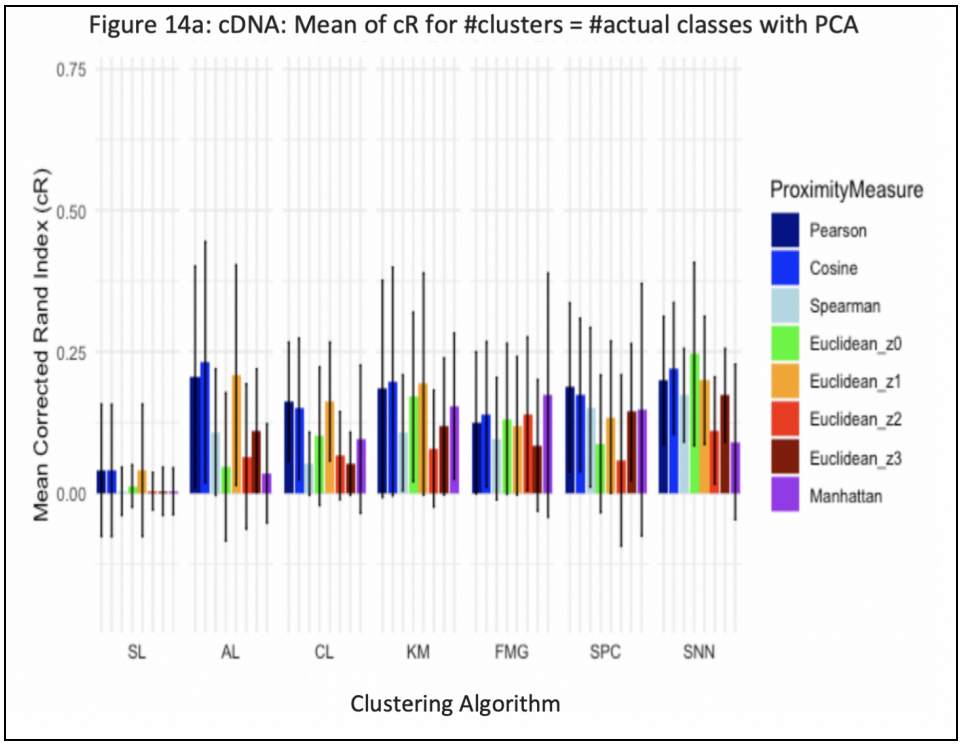
cDNA: Mean of cR: *The mean of the cR for k equals number of actual classes (a) and mean of cR for best partition (b) is displayed. The missing bars indicate that the proximity measures were not compatible with the algorithm.*
— Ekene

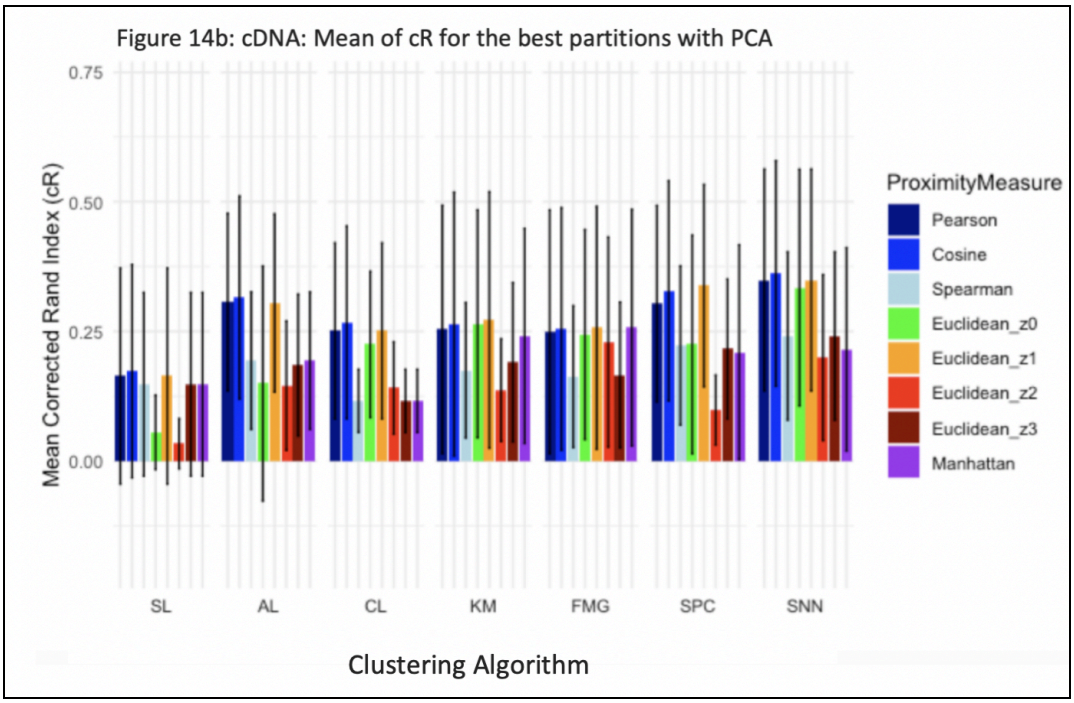
Figure 4: In the published figure, the hierarchical clustering methods have the greatest mean differences, and k-means and FMG have the lowest differences. There are similar results in the replicated figure, except SPC with Euclidean_z2 also has a high mean difference and SNN has the lowest mean difference.



cDNA: Mean Difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number clusters from the best partition for cDNA datasets.* — Ekene

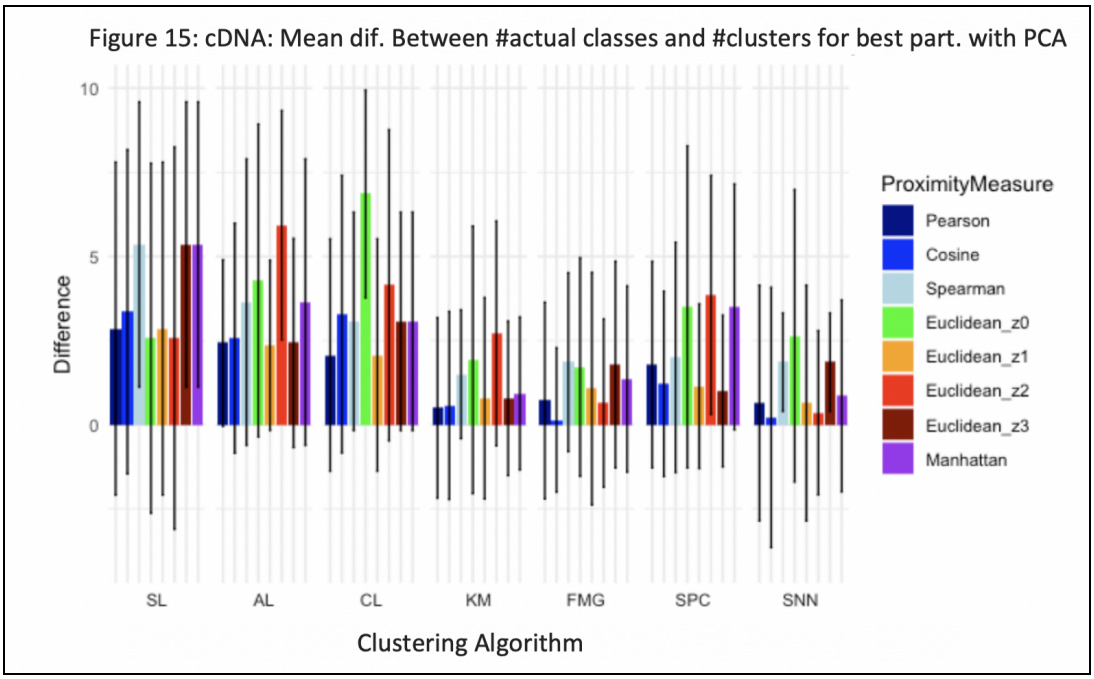
Figure 14a and 14b: The replication of Figure 2a with PCA, Figure 14, generally maintains the same structure as the non-PCA Figure 2a, but the cR index is lower which means the performance of the clustering algorithm worsened with PCA. The same findings for Figure 14a hold true for Figure 14b.





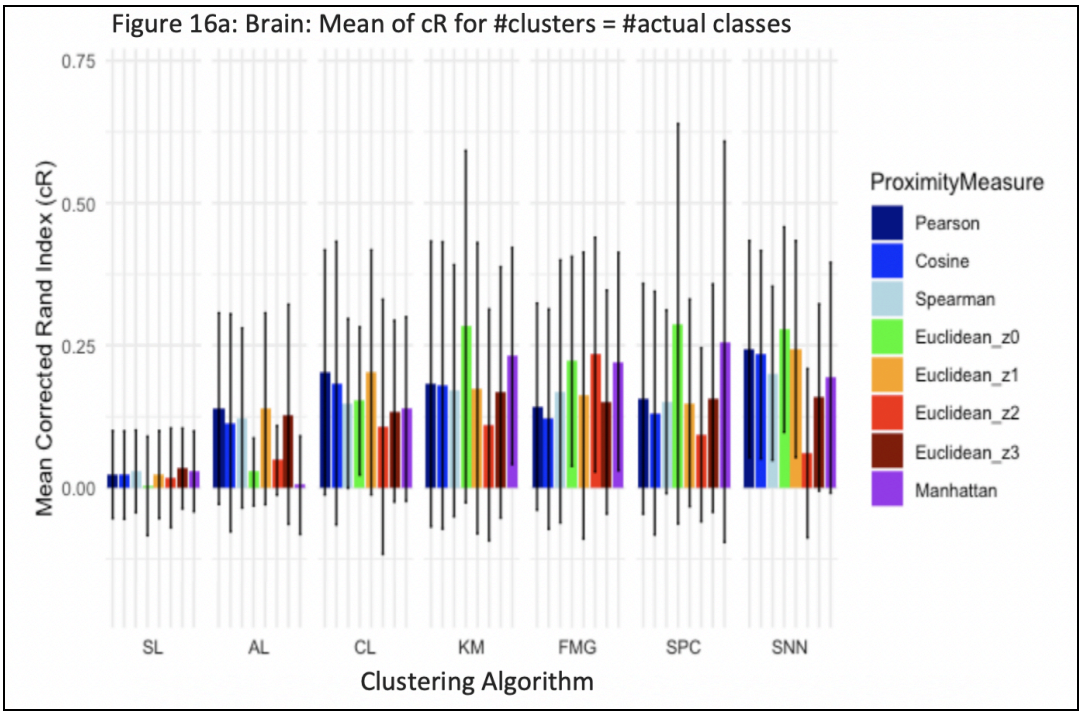
cDNA with PCA: Mean of cR: *The mean of the cR for k equals number of actual classes (a) and mean of cR for best partition (b) is displayed. The missing bars indicate that the proximity measures were not compatible with the algorithm.* — Ekene

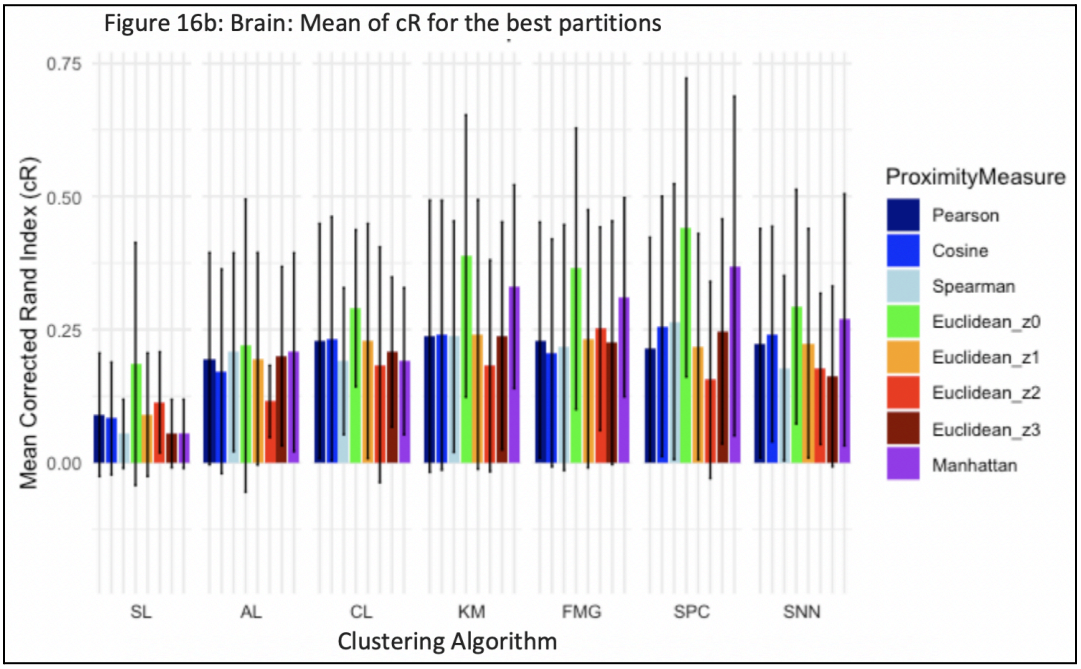
Figure 15: Figure 15 displays similar results as Figure 4–hierarchical clustering methods have greater differences–however, Figure 15 also shows more differences between proximity measures in the same clustering algorithm. For example, only Euclidean_z2 of SPC without PCA had a high difference while the rest were low. However, SPC with PCA shows three bars with higher difference and the rest. A similar phenomenon can also be observed with the hierarchical clustering methods.



cDNA with PCA: Mean Difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number clusters from the best partition for cDNA datasets with PCA.* — Ekene

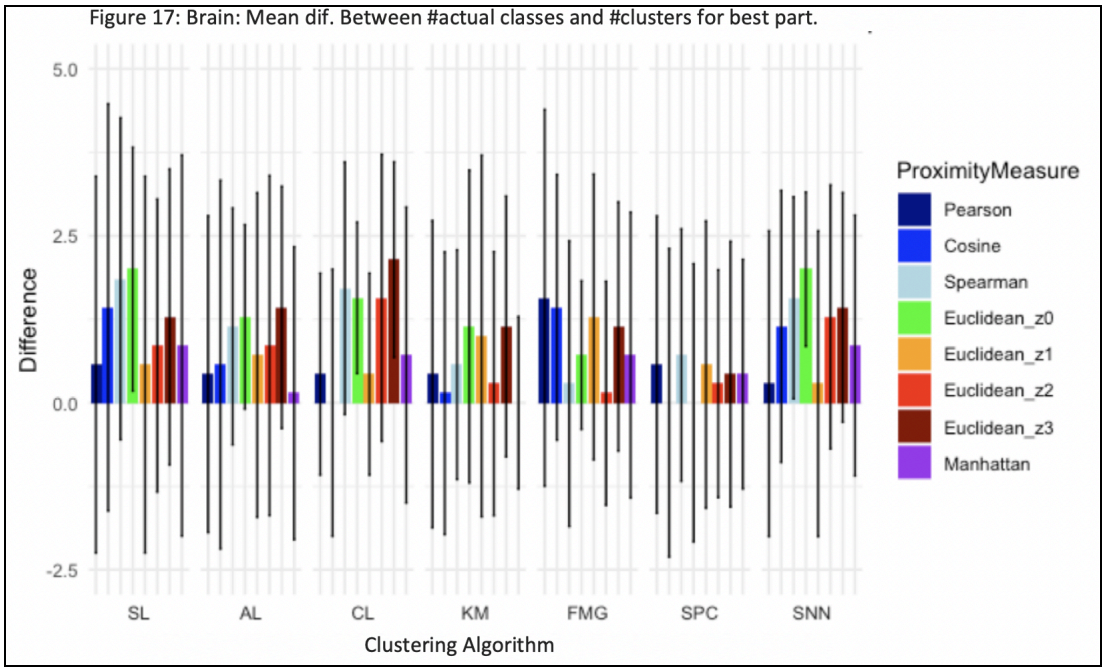
Figure 16a and 16b: These figures were created with Affymetrix and cDNABrain tissue datasets. For Figure 16a no one clustering method significantly performs best overall. However, k-means, SPC, and Shared Nearest Neighbor contain the bars with the greatest cR values, and single-linkage performs significantly worse than all other methods. Also, Euclidean_z0 gives the highest cR values for nearly every clustering method. Figure 16b shows the same trends as found in Figure 16a, except only k-means and SPC contain the bars with the greatest cR values.





Brain: Mean Difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number clusters from the best partition for Brain Tissue datasets.* — Ekene

Figure 17: The differences for all clustering methods are much more similar than observed in Figure 4. However, the hierarchical clustering methods still appear to produce the greatest overall mean difference. Shared Nearest Neighbor also has a noticeably high mean difference too.



Brain: Mean Difference between actual and best clusters: *This displays the differences in the actual number of clusters versus the number clusters from the best partition for Brain Tissue datasets.* — Ekene

Liz's Contribution:

Figures 5a, 5b, and 6 show the replication of Figures 5a, 5b, and 6 from the original paper, which employ a cDNA dataset (**Alizadeh-2000-v2**) containing ***blood*** tissue samples. **Figure 5a** was replicated using hierarchical clustering and matches the published figure. Both generally classify blood tissue samples into 3 clusters. Additionally, both contain a single wrongly assigned DLBCL sample and one of the clusters has FL and CLL samples combined.

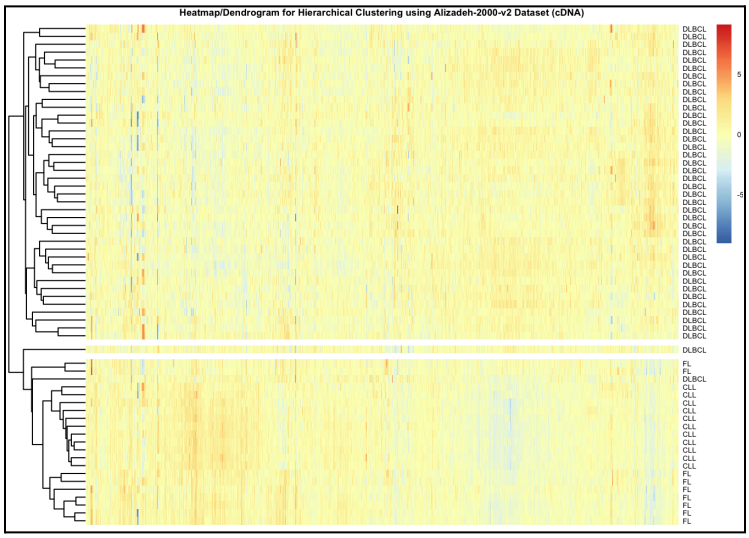


Figure 5a: Hierarchical clustering for Alizadeh-2000-v2 dataset (cDNA) — Liz

Similarly, **Figure 5b** was replicated using k-means clustering, and the same results as the original figure were obtained. Both published and replicated heatmaps show the tissue samples in 3 clusters. If we ignore the single misclassified DLBCL sample, each cluster contains one distinct cancer type, a significant improvement from hierarchical clustering.

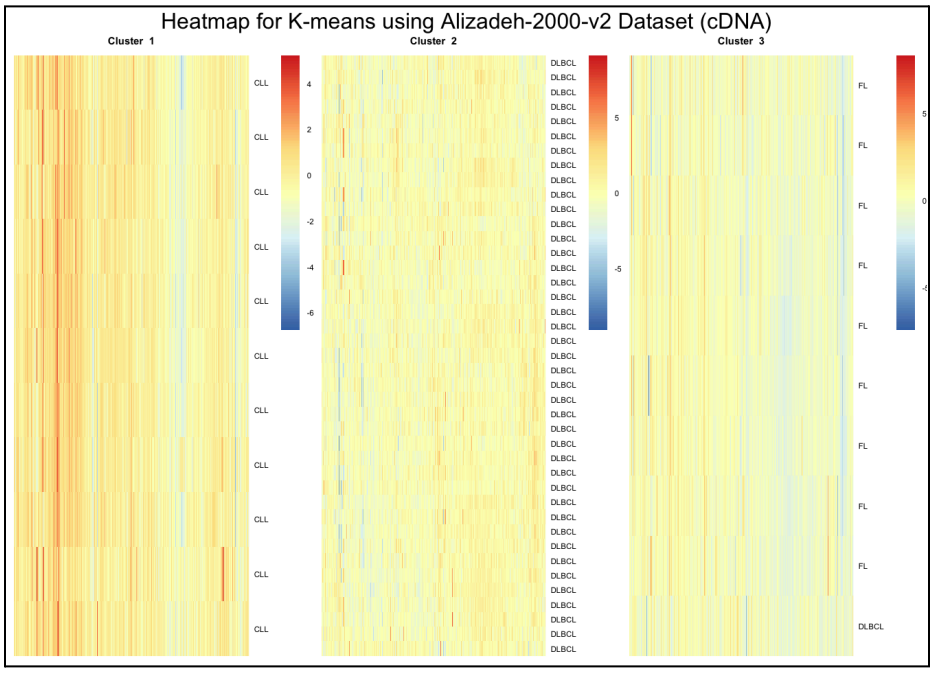


Figure 5b: K-means clustering for Alizadeh-2000-v2 dataset (cDNA) — Liz

To replicate **Figure 6**, PCA analysis was performed and plotted using the top two largest components. The replicated figure shows the same results as the published figure, where the DLBCL, CLL and FL samples are grouped into three distinct clusters.

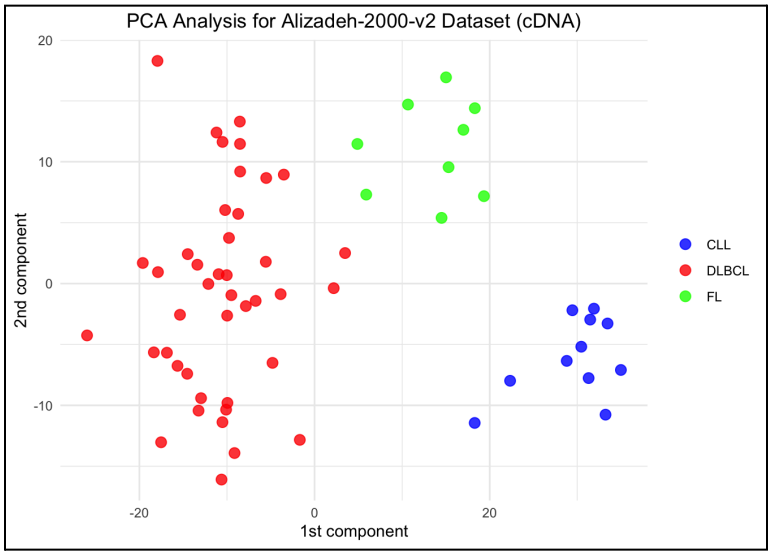


Figure 6: PCA plot for Alizadeh-2000-v2 dataset (cDNA) — Liz

Figures 7 and 8 are an extension of Figure 5a and Figure 5b respectively, where hierarchical and k-means clustering are performed on an Affymetrix dataset (**Armstrong-2000-v2**) instead of a cDNA dataset. This dataset also contains ***blood*** tissue samples. Compared to published Figure 5a, **Figure 7** shows the same patterns except that it has an additional misclassified sample. Similarly, k-means in **Figure 8** shows the same trends as the original Figure 5b, but also has an additional misclassified sample.

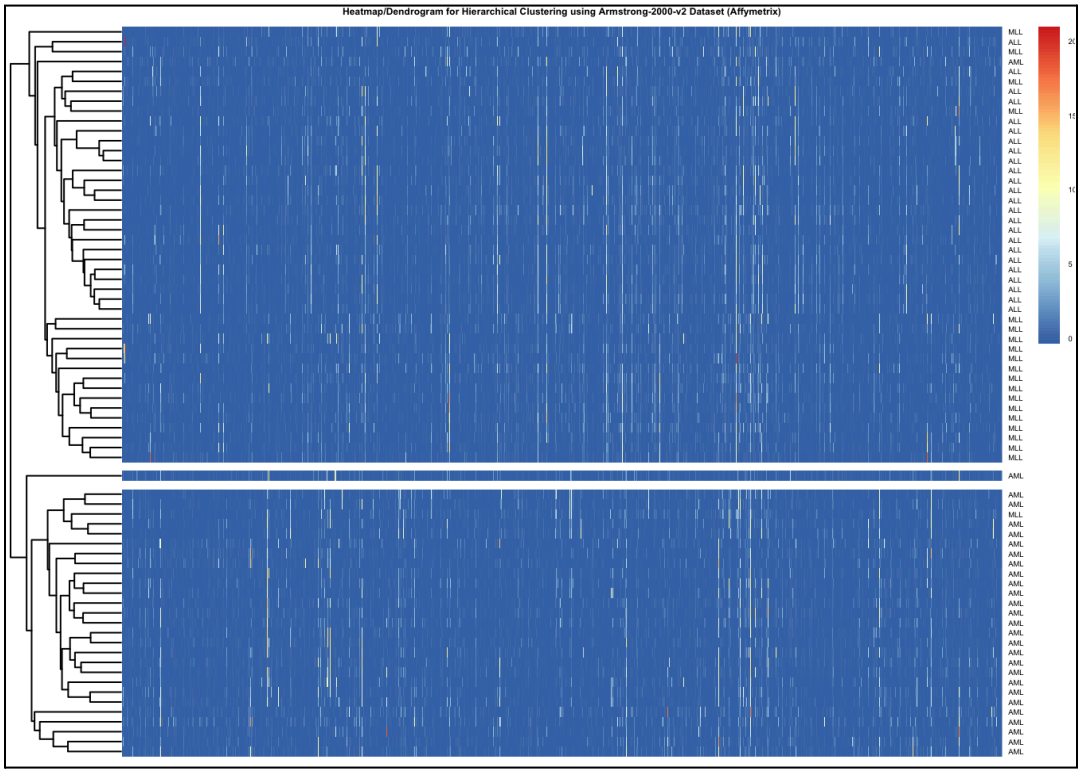


Figure 7: Extension of Figure 5a – hierarchical clustering for Armstrong-2000-v2 dataset (Affymetrix) — Liz

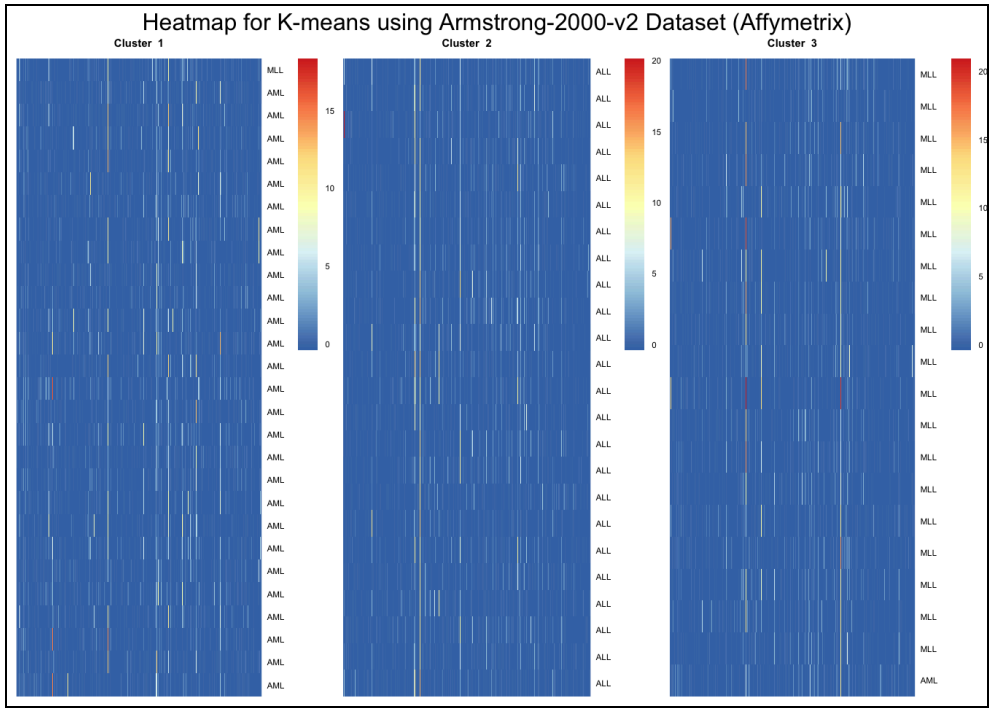


Figure 8: Extension of Figure 5b – k-means for Armstrong-2000-v2 dataset (Affymetrix) — Liz

In contrast to the PCA plot for the cDNA dataset which illustrated three distinct clusters, **Figure 9** has more ambiguous clusters, since samples from different types mix or are located near each other.

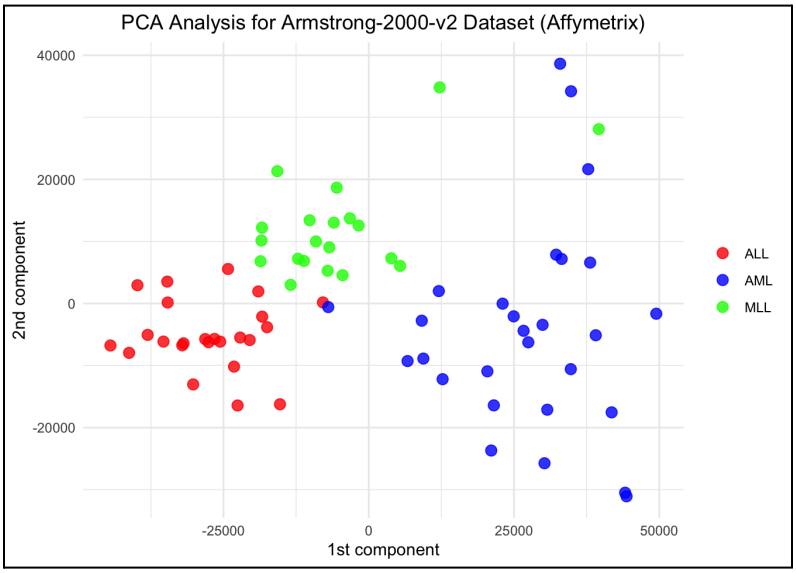


Figure 9: Extension of Figure 6: PCA plot for Armstrong-2000-v2 dataset (Affymetrix) — Liz

As an additional extension of Figure 5, a comparison between k-means and FMG (Finite Mixture Gaussian) – the clustering algorithms with the highest performance reported in the original study [1] – was executed using an Affymetrix and a cDNA dataset containing ***brain*** tissue samples.

Figure 10 and 11 demonstrate that both clustering algorithms were unsuccessful in classifying the samples, as both resulted in clusters that contain a mixture of samples from different classes. The Affymetrix dataset, **Nutt-2003-v1**, has 4 classes while the cDNA dataset, **Bredel-2005,** has 3 classes.

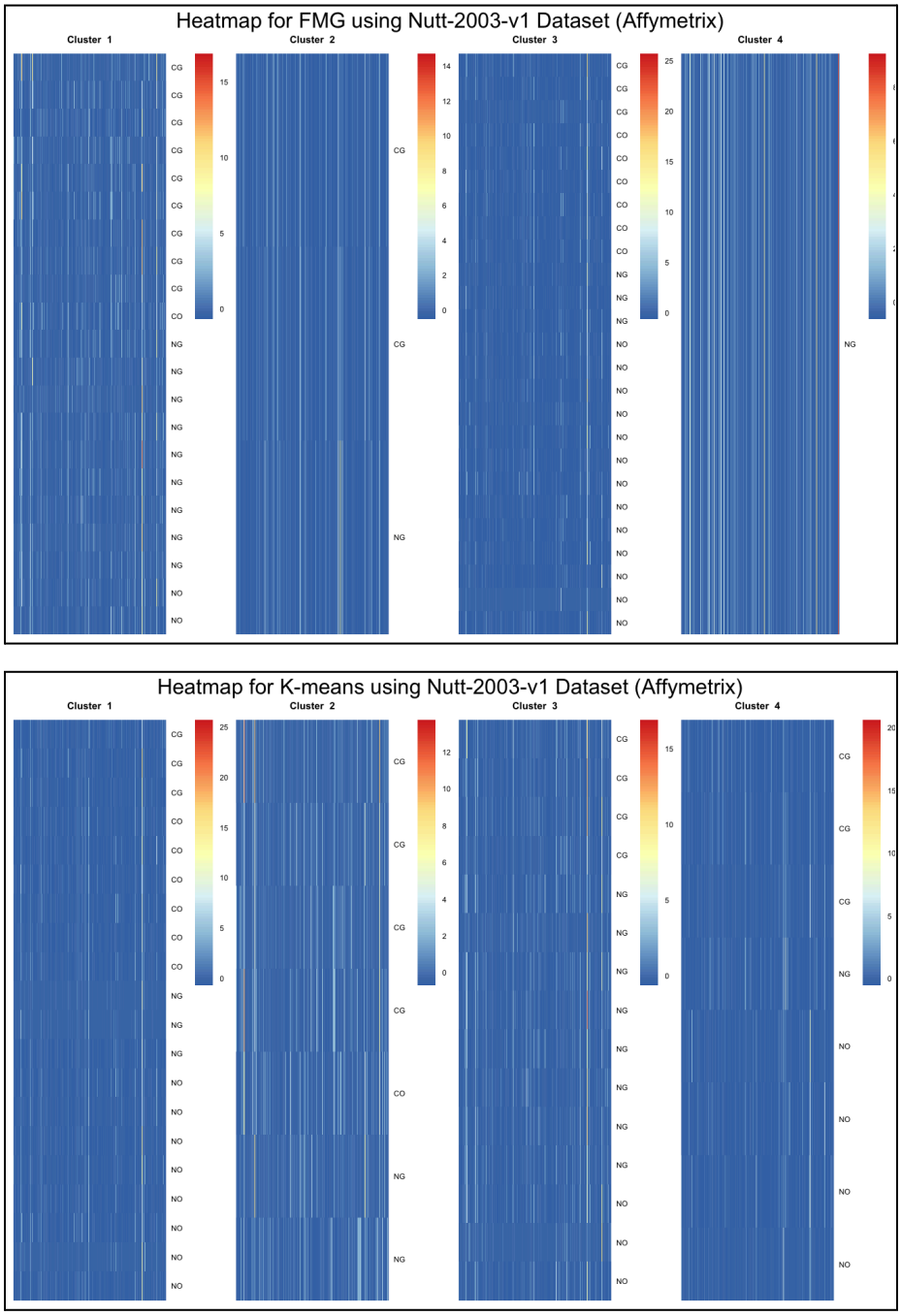


Figure 10: FMG and k-means clustering with Nutt-2003-v1 dataset (Affymetrix) — Liz

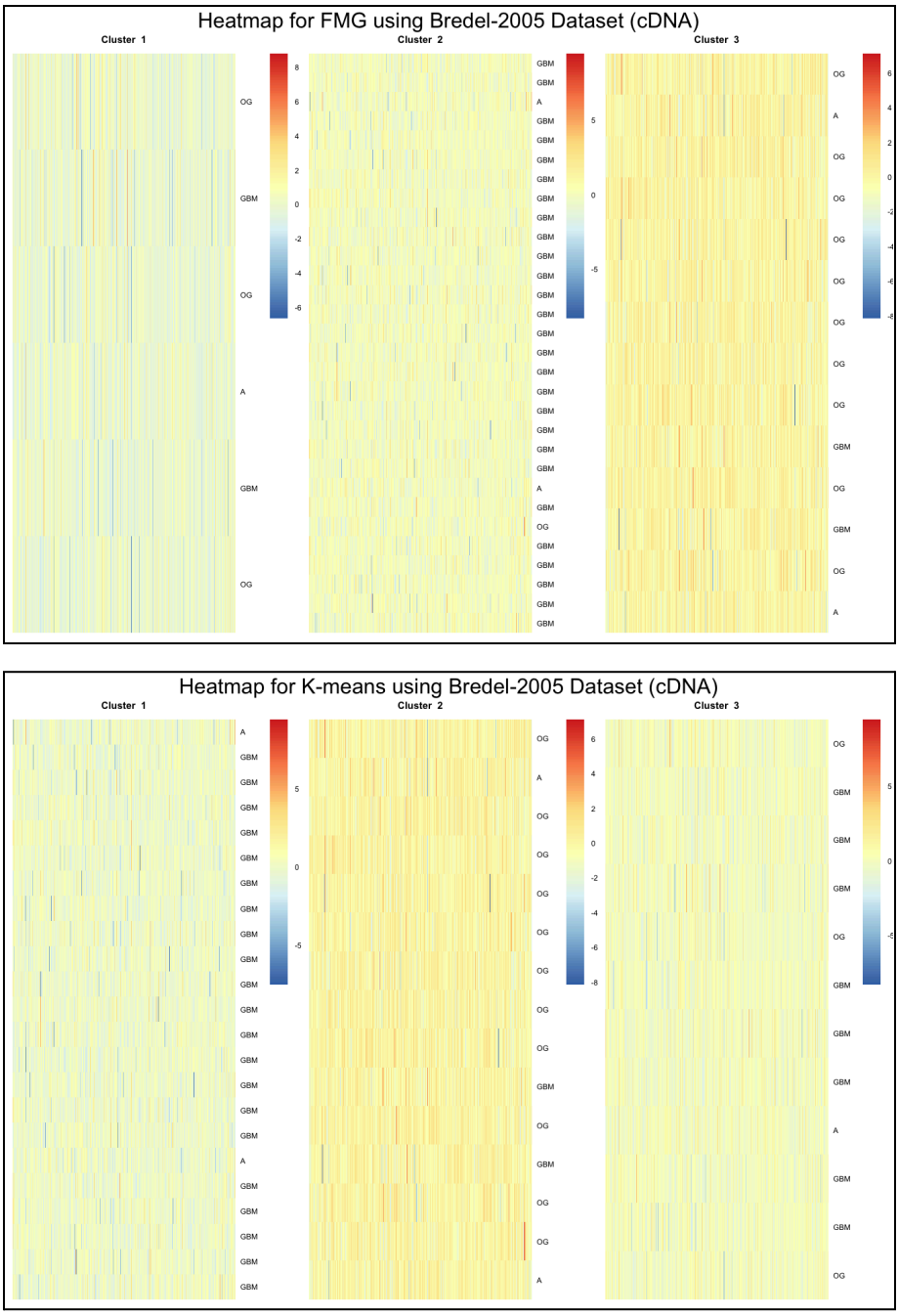


Figure 11: FMG and k-means with Bredel-2005 dataset (cDNA) — Liz

PCA analysis was performed on the Affymetrix and cDNA datasets from **Figures 10 and 11** and plotted using the two largest components. The results are shown in **Figures 12 and 13**. PCA analysis fails to recognize the 4 cancer subtypes in the Affymetrix dataset and the 3 cancer subtypes in the cDNA dataset, as samples from different types mix and are scattered around.

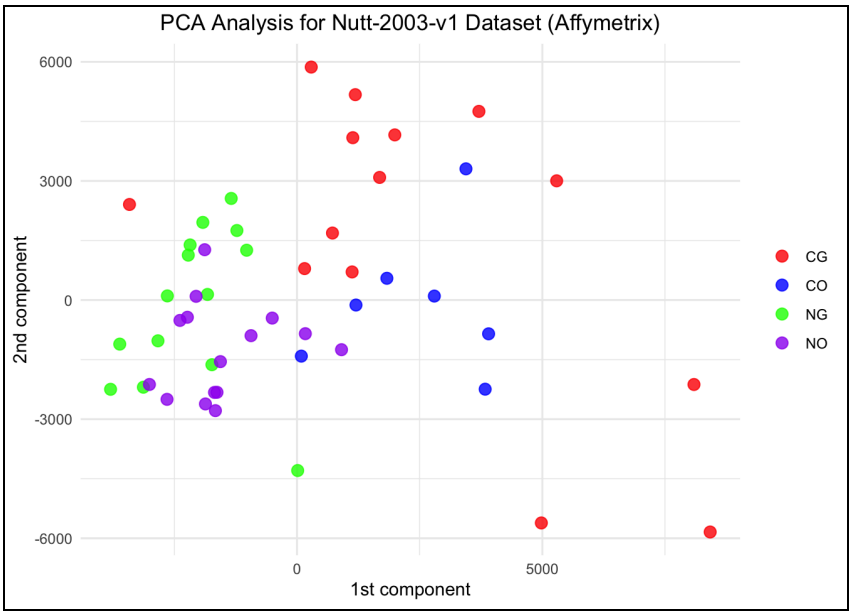


Figure 12: PCA Analysis with Nutt-2003-v1 dataset (Affymetrix) — Liz

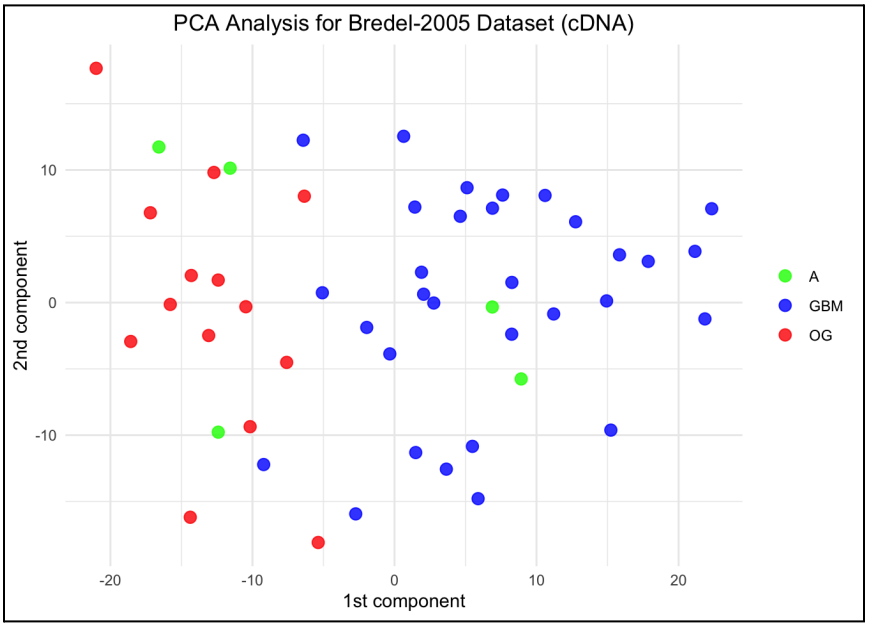


Figure 13: PCA Analysis with Bredel-2005 dataset (cDNA) — Liz

Discussion

Victoria: Figures 1, 3 shows that the best algorithm is k-means, while it was FMG in the paper, and my mean cR index was lower. Thus, my figures failed to recapitulate the results. The difference may result from the algorithms' implementation. For instance, there are many implementations of SNN and

values of function parameter impacts result, like in k-means. Comparing Figures 1, 3 and 2, 4, Manhattan distance performed better in Affymetrix than cDNA datasets, likely due to differences in Affymetrix and cDNA technologies. Low mean cR index in Figures 14a, 14b, 18 and 19 reveals that PCA didn't improve performance, which can be explained by loss of information, revealing that clustering cancer subtypes may need informative data. Figure 15 shows differences between measures in the same algorithm, likely because PCA removed important information, causing less predictable results. Lastly, Figures 16, 17 and 20, 21 shows that the algorithms performed better on blood than brain tissue datasets, thus different tissue types play a role. These results are beneficial for future work in determining algorithms for gene expression datasets. Future work includes trying out other clustering algorithms, hybrid methods, and new microarray technologies.

Ekene: The recapitulation of the results from Figures 2 and 4 was unsuccessful. The cR values were much lower, and the order of the best clustering methods differed in the replicated figures. For cDNA datasets, when the number of clusters should match the number of classes, FMG with Pearson's correlation or SNN with Euclidean_z0 perform best and single linkage is worst (Figure 2a). For the best partitions in cDNA datasets, SPC and SNN with Cosine or Euclidean_z1 are best (Figure 2b). Hierarchical clustering methods and Euclidean_z1 distance metric produce the most significant differences in Figure 4, so these methods/metrics are not efficient at capturing the true number of classes to produce the best cR. For Brain tissue datasets, when the number of clusters should match the number of classes, k-means, SPC, or SNN with Euclidean_z0 are the best clustering methods, and single linkage is the worst (Figure 16a). For the best partitions in Brain tissue datasets, SPC Euclidean_z0 is the best clustering method (Figure 16b). Single linkage, complete linkage, and SNN produce the greatest difference in Figure 17, so we can apply the same conclusion as we did with Figure 4. The opposite is true for SPC in Figure 17. Finally, potential pitfalls arise from the clustering method implementation. For example, there are different SNN methods, and we chose the best one, but we can get different findings from the original study depending on our implementation.

Liz: Regarding **Figure 5a, 5b and 6**, as the original study [1] mentions, it is important to note that without prior knowledge of the distinction between FL and CLL samples, the use of hierarchical clustering alone would not indicate the existence of these two classes. In contrast, the distinction between FL and CLL samples is detected by k-means and clearly visualized with PCA analysis. **Figures 7 and 8** illustrate the results from performing hierarchical clustering and k-means clustering on an Affymetrix dataset, where there is an additional-misclassified sample compared to the original published Figure. This suggests the performance of hierarchical clustering and k-means in classifying the samples is slightly worse for Affymetrix than cDNA. These results are supported by the PCA analysis of **Figure 9 vs. Figure 6** and **Figure 12 vs. 13**, as the ability to correctly group samples into compact clusters diminishes

for the Affymetrix dataset compared to the cDNA dataset. Comparing **Figures 5-9** (using blood tissue datasets) and **Figures 10-13** (using brain tissue datasets), it can be observed that the classification quality on the datasets containing blood tissue samples was better than that on datasets containing brain tissue samples. This suggests that the tissue type of samples matters for classification.

Future research can investigate the reason classification performance of the blood and brain samples is worse for the Affymetrix than the cDNA datasets. Additionally, it would be interesting to investigate why there was better classification for the blood tissue datasets in **Figures 5-9** compared to the brain tissue datasets in **Figures 10-13**.

Conclusion

Based on **Figures 1 and 4**, we can conclude that k-means and FMG were the top two clustering algorithms for Affymetrix datasets, which aligned with the paper's results. In general, Pearson and Cosine performed best out of all the proximity measures. From **Figures 2a, 2b,** and **4**, we can conclude that SNN is the best clustering method for cDNA datasets. In the case of Figure 2a, Euclidean_z0 is the best proximity measure for SNN while for Figure 2b Pearson, Cosine, or Euclidean_z1 would be best. From **Figures 14, 15, 18, and 19**, we can conclude that using PCA on the gene expression datasets is not ideal because we do not achieve as high of cR as we do without PCA. Therefore, the dimensionality reduction that we achieve with PCA must be removed from details/components that are useful for clustering. From **Figures 16a, 16b,** and **17**, we can conclude that SPC is the best clustering method for Brain Tissue Affymetrix and cDNA datasets. We also conclude that Euclidean_z0 is the best proximity metric for SPC. From Figures 20 and 21, we can conclude that spectral clustering and SNN were the best clustering algorithms for blood tissue datasets. Pearson and Cosine still performed the best out of all the proximity measures, which means that it might be best to use these proximity measures for cancer gene expression data. From **Figures 5-9,** we conclude better classification performance by k-means clustering and PCA compared to hierarchical clustering when applied to a cDNA dataset (Alizadeh-2000-v2) or an Affymetrix dataset (Armstrong-2000-v2). From **Figures 5-13**, we observe better classification performance on cDNA datasets than Affymetrix, and stronger classification performance on blood tissue datasets over brain tissue datasets in general.

Methods

<u>Victoria's Contribution:</u>

The 7 clustering algorithms (single linkage, complete linkage, average linkage, k-means, FMG, SPC, and SNN) were first applied to the Affymetrix datasets, with the different proximity measures

(Pearson, Cosine, Spearman, original Euclidean distance, standardized Euclidean distance, scaled Euclidean distance, and ranked Euclidean distance) with an additional proximity measure of Manhattan distance.

The proximity measures were applied to the datasets first to create a distance matrix then passed into the clustering algorithms, where the value of k was equal to the actual number of classes in Figures 1a, 18a, and 20a. In Figures 1b, 18b, and 20b, the value of k was the best partition. Then, the corrected Rand index was calculated for each dataset and the mean was taken across each algorithm. The bar plots in these figures were created using the mean cR index and using standard deviation to calculate the error bars. Next, the difference between the actual number of classes and the best partition was calculated to generate the bar plots for Figures 3, 19, and 21. For Figures 18 and 19, PCA was applied to the Affymetrix datasets before using the clustering algorithms and the proximity measures. Lastly, the same procedure was applied for Figures 20 and 21, but with blood tissue datasets instead.

Ekene's Contribution:

In my replication of **Figures 2a, 2b, 4**, the same clustering algorithms and distance metrics were used, but an additional distance metric of Manhattan distance was used. I replicated **Figure 2a** by first computing a distance matrix for each type of distance proximity measure. I computed the clusters for each combination of distance measures and clustering methods (there is a separate file for every clustering method's results), and used that information to compute the adjusted Rand Index (this is the same as the cR). I repeated this process for each cDNA dataset. I then found which datasets for each proximity measure had the same number of clusters as the number of actual classes, and used the cR of those clusters to calculate the mean cR and the error values. The error values were calculated using standard deviation. As I did in Figure 2a, I replicated **Figure 2b** by first computing a distance matrix for each type of distance proximity measure. I also computed the clusters for each combination of distance measures and clustering methods (there is a separate file for every clustering method's results), and used that information to compute the adjusted Rand Index. I repeated this process for each cDNA dataset. And for each cDNA dataset. I repeated this entire process for all values of k where k is the suggested number of clusters. I then found the greatest cR values for all 14 datasets. I computed the mean of the cR for the best partitions for each proximity measure by taking the average of the best cR values for each dataset. I also used the best cR values to calculate the error. I replicated **Figure 4** by repeating the same steps as Figure 2b. However, I collected the k values that produced the greatest cR values for all 14 datasets, and subtracted the true k-value (actual number of classes) from best k-values. The absolute value of this gave the difference. I also used the differences to calculate the error. For **Figures 14a, 14b,** and **15**, I first applied PCA to the datasets, then I used the same exact procedure as was used for Figures 2a, 2b, and 4.

For **Figures 16a, 16b,** and **17** I used the exact same procedure as was used for Figures 2a, 2b, and 4; however, the inputs came from the Brain tissue of both Affymetrix and cDNA datasets.

Liz's contribution:

To produce **Figures 5a and 7,** hierarchical clustering was conducted and visualized using a heatmap with dendrogram using the pheatmap library. I calculated Pearson's correlation distance by computing the correlation matrix with "Pearson" as the method and subtracting 1 by this correlation matrix. To match the orientation of the original paper's figure, I transpose the corresponding dataframe (such that samples are along rows and genes are along columns), before using it in the pheatmap function. Within this function, I use the computed Pearson's correlation distance matrix, set 'cutree_rows' = 3 to cut the tree into 3 clusters along rows, set clustering_method = "average" to use average linkage, cluster_rows = TRUE since we are classifying samples, scale = none, and show_rownames = TRUE to display sample labels.

To generate the k-means clustering illustrations of **Figures 5b, 8, 10, and 11**, I performed k-means clustering with the 'kmeans' R function. To generate the FMG clustering illustrations of **Figures 10 and 11**, I employ the 'mclust' library. Then, I scale the data to standardize gene expression profiles for each sample, and then transpose the data frame such that samples are along rows and genes are along columns. Next, I pass this dataframe to the respective classifier function with an argument of k corresponding to the number of true classes of the dataset being analyzed. Using the obtained cluster assignments, I reorder the samples (rows) of the dataframe. Thereafter, I also order the cluster assignments' order to match the new order of the dataframe. Finally, I generate a heatmap with all clustering parameters set to FALSE (since we already have clustered the rows) and show_rownames = TRUE to display sample labels.

To generate **Figures 6, 9, 12, and 13**, I first transpose the data frame such that samples are along rows and genes are along columns. Then I pass in the transposed dataframe to the 'prcomp' R function to perform PCA. Next, I flip the sign for the second component values to match the original paper's figure's orientation. Finally, using the 'ggplot2' library, I make a scatterplot of the top two largest principal components.

Author contributions

Liz was responsible for **Figures 5-13** in the results section, along with the respective procedure in the methods section and the analysis for each of these figures in the discussion and conclusion sections.

Ekene was responsible for **Figures 2, 4** and **14-17** in the results section, along with the respective procedure in the methods section and the analysis for each of these figures in the discussion and conclusion sections.

Victoria was responsible for **Figures 1, 3** and **18-21** in the results section, along with the respective procedure in the methods section and the analysis for each of these figures in the discussion and conclusion sections.

Acknowledgements

We would like to thank the professors and the teaching assistants for guiding and providing us feedback throughout the project.

References

1. de Souto, M.C., Costa, I.G., de Araujo, D.S. *et al.* Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **9**, 497 (2008). https://doi.org/10.1186/1471-2105-9-497
2. Schliep, Alexander. "Supplementary Material: Clustering Cancer Gene Expression Data: A Comparative Study." *Schliep Lab*, schlieplab.org/Supplements/CompCancer/