High Dimensional Statistics IV MATH4287 - Mini Project Report

Does dimensionality reduction improve clustering performance in high-dimensional gene expression data?

qvns53@durham.ac.uk, Durham University

2025-12-19

1 Introduction

High-dimensional biological datasets pose substantial challenges for classical clustering methods. In particular, gene expression microarray data typically exhibit a $p >> n$ structure, where the number of measured variables far exceeds the number of available observations. In such settings, Euclidean distances tend to concentrate, reducing their discriminative power and potentially degrading the performance of distance-based clustering algorithms.

A common strategy for addressing these issues is to apply dimensionality reduction prior to clustering. Principal Component Analysis (PCA) is frequently used to project the data onto a lower-dimensional subspace that captures most of the total variance. However, directions of maximal variance do not necessarily correspond to directions that best separate clusters, and dimensionality reduction may therefore fail to improve clustering performance.

This mini-project investigates whether dimensionality reduction improves clustering performance on a real high-dimensional dataset: the Human Tumor Microarray dataset. We compare clustering results obtained using raw data, PCA-reduced data, and sparse PCA-reduced data, focusing on both clustering accuracy and interpretability.

2 Objectives & Questions

The objectives of this project are:

- To evaluate whether dimensionality reduction improves clustering performance in a high-dimensional, low-sample-size setting.
- To compare standard PCA and sparse PCA as preprocessing steps for clustering.
- To assess the effect of dimensionality reduction on different clustering algorithms, focusing the evaluation on objective metrics as well as interpretability.

The primary research question is:

- Does dimensionality reduction improve clustering performance, as measured by the Adjusted Rand Index, on the Human Tumor Microarray dataset?

Secondary questions include:

- Does sparse PCA offer advantages over standard PCA for clustering?
- How does dimensionality reduction affect different hierarchical linkage methods?

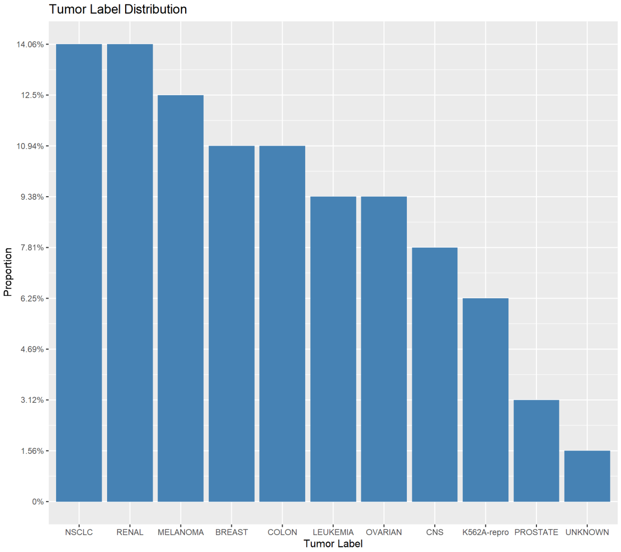
3 Literature Review

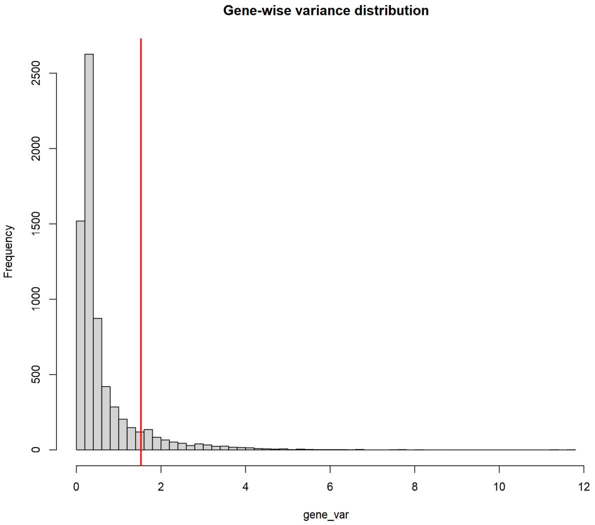
This report references various literature sources [1], [2], [3], [4], [5], [6], as well as the course materials.

4 Methods

4.1 Dataset Description & Exploratory Analysis

The Human Tumor Microarray dataset consists of gene expression measurements for $n = 64$ tumour samples across $p = 6830$ genes. Each sample is associated with a known tumour type label, which is used only for external evaluation of clustering performance. The class distribution is imbalanced, with some tumour types appearing much more frequently than others (see the left figure below). Note the 'PROSTATE' and 'UNKNOWN' classes are the least common, so they would likely be absorbed by the more frequent classes during the running of a clustering algorithm.





All gene expression variables were centred by subtracting their sample means, as this is standard practice. Scaling was not applied, as the analysis focuses on variance structure and relative differences between samples.

Exploratory analysis revealed that many genes exhibit very low variance across samples, while a small subset shows substantially higher variability (see the right figure above). This suggests that much of the signal may lie in a lower-dimensional subspace, motivating the use of PCA and sparse PCA.

4.2 Clustering Algorithms

Two clustering approaches were considered:

- k-means clustering, using Euclidean distance.
  - $d(X_i, X_j) = ||X_i - X_j||_2^2 = \sum_{l=1}^{p} (X_il - X_jl)^2$
- Hierarchical (agglomerative) clustering, with single, average, and complete linkage.
  - single: $d_{SL}(G, H) = min_{i \in G, i^\star \in H} d_{ii^\star}$
  - complete: $d_{CL}(G, H) = max_{i \in G, i^\star \in H} d_{ii^\star}$
  - group average: $d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i^\star \in H} d_{ii^\star}$

To allow fair comparison across methods and dimensionality reductions, the number of clusters was fixed at $k = 5$. This choice was guided by elbow plots and Adjusted Rand Index (ARI) values obtained from k-means clustering on the raw data. See Section "5.1 Choosing the k value' for more details on this.

4.3 Dimensionality Reduction

PCA was applied to the centred data using singular value decomposition. Typically the number of principal components is chosen to pick a 'cut-off' at a component that has significantly more variance than the one immediately after. This is practically done by using a scree plot visualisation and applying the 'elbow method'. An alternative to that visual method is to use the cumulative proportion of variance explained. Thresholds of 75%, 90%, and 95% variance explained were considered, corresponding to substantial reductions in dimensionality.

Sparse PCA was also applied, imposing sparsity on the loading vectors to encourage selection of a small subset of genes contributing to each component. This approach aims to improve interpretability while retaining the main variance structure.

4.4 Evaluation Metrics

To evaluate clustering methods, one can try to visualise the clusters and check for cluster separation across pairs of dimensions. That method is clearly visual and therefore subjective, but it can still be very useful if you pick a new pair of dimensions that are likely to lead to good performance. To do this, one can use PCA again, (completely separately to using it to reduce dimensionality before clustering) as a visualisation aid; then plot the clusters along the first 2 principal components and check for separation.

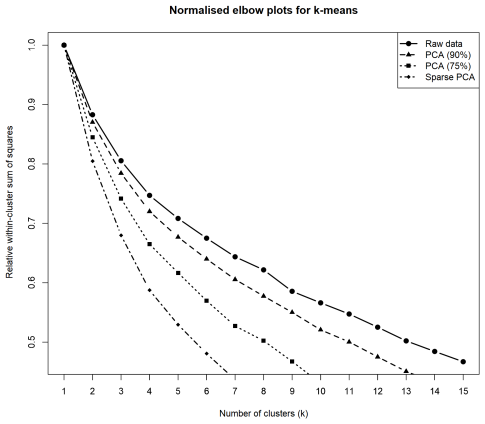
Clustering performance was further evaluated using the Adjusted Rand Index (ARI), which measures agreement between the clustering assignment and the true tumour labels while correcting for chance.

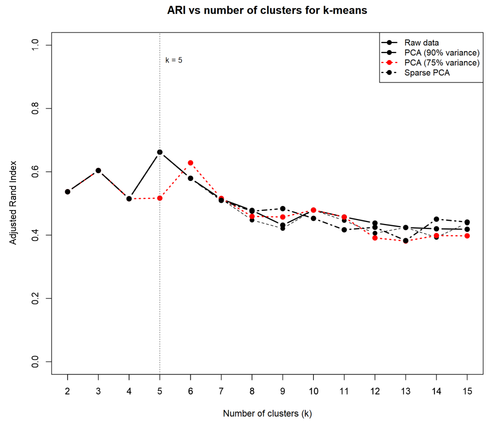
5 Results & Findings

5.1 Choosing the k value

To pick the optimal k number of clusters to use throughout the whole analysis, we used a scree plot of number of clusters (k) against the within-cluster sum of squares (WSS). To pick a suitable k, one should find the "elbow" of this plot, but as is visible on the plot, there is not a clear value of k for this data.

As an alternative, we used the ARI plotted against k, and looked for peaks. The most optimal choice from these plots is $k = 5$ which was used for all methods.

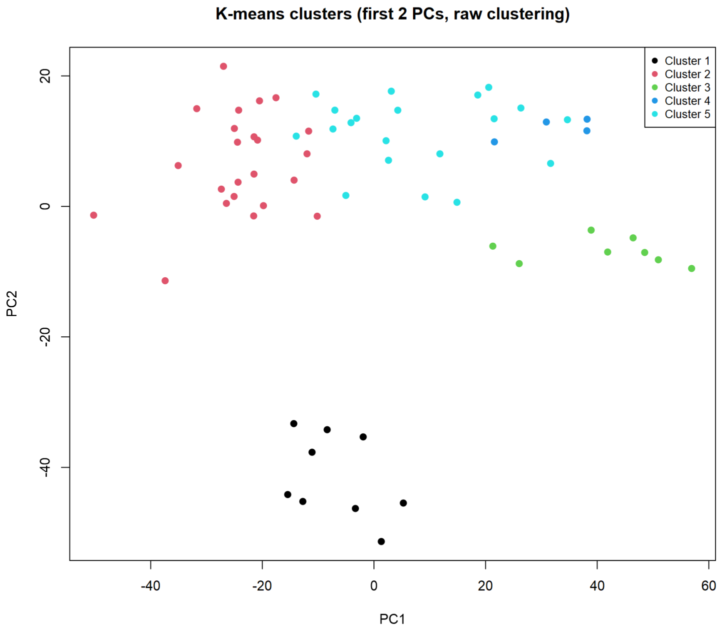




5.2 Baseline clustering

Applying k-means clustering directly to the full high-dimensional dataset yielded an ARI of approximately 0.66, which was the highest observed across all methods considered. Hierarchical clustering performed less well overall, with complete linkage outperforming single and average linkage.

Despite the high dimensionality, clustering on the raw data retained meaningful structure, suggesting that the signal is not completely overwhelmed by noise.

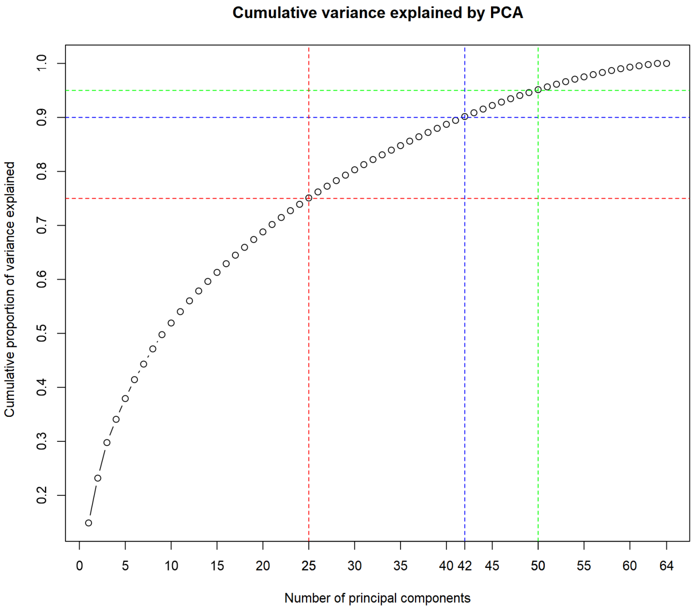


5.3 PCA-based clustering

The cumulative variance explained by PCA is shown below. Approximately 90% of the total variance is captured by the first 42 principal components, representing a substantial reduction from the original 6830-dimensional space.

However, applying k-means clustering to PCA-reduced data with 90% or 95% variance retained produced clustering results identical to those obtained on the raw data for $k = 5$, yielding the same ARI. When only 75% of the variance was retained, the clustering structure changed, but no improvement in ARI was observed.

The right figure of Section 5.1 shows ARI as a function of the number of clusters for raw data, PCA-reduced data, and sparse PCA. Across most values of k, dimensionality reduction did not improve clustering accuracy.

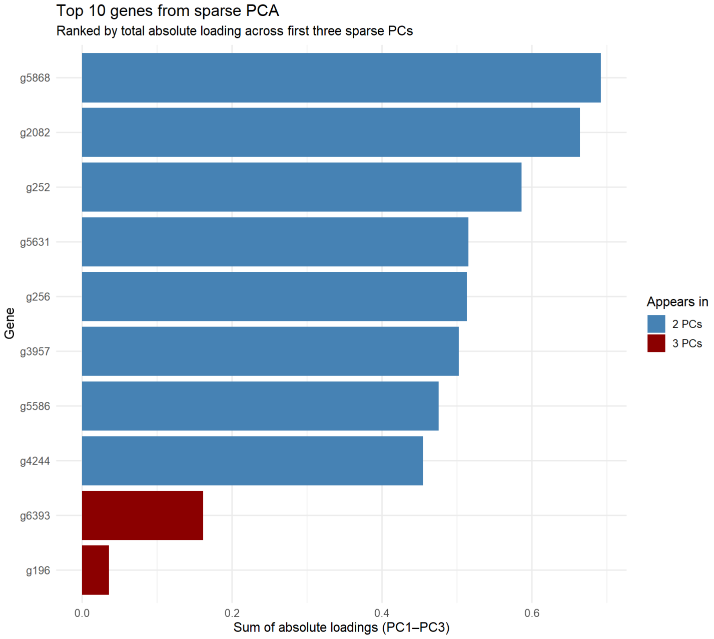


5.4 Sparse-PCA based clustering

Sparse PCA produced clustering results similar to those obtained using standard PCA, with no improvement in ARI for k-means clustering. However, sparse PCA yielded loading vectors involving a small subset of genes, facilitating interpretation of the principal components.

An analysis of sparse PCA loadings revealed that certain genes appeared repeatedly across the first few components, suggesting potential biological relevance. This interpretability advantage distinguishes sparse PCA from standard PCA, despite similar clustering performance.

The most impactful genes are shown in the figure below.



6 Conclusions

- Clustering on the raw high-dimensional data achieved the highest Adjusted Rand Index.
- PCA-based dimensionality reduction did not improve clustering performance, even with substantial reductions in dimensionality.
- Aggressive dimensionality reduction (75% variance) altered clustering structure but did not enhance accuracy.
- Sparse PCA did not improve clustering accuracy but provided improved interpretability through sparse gene loadings.
- These results highlight that dimensionality reduction does not necessarily improve clustering performance in high-dimensional biological data.

Please see https://github.com/Shulux/HDS/tree/main for the full code used to produce the plots in this report.

7 References

[1] K. Y. Yeung and W. L. Ruzzo, "An empirical study on Principal Component Analysis for clustering gene expression data," 2001.

[2] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/BF01908075.

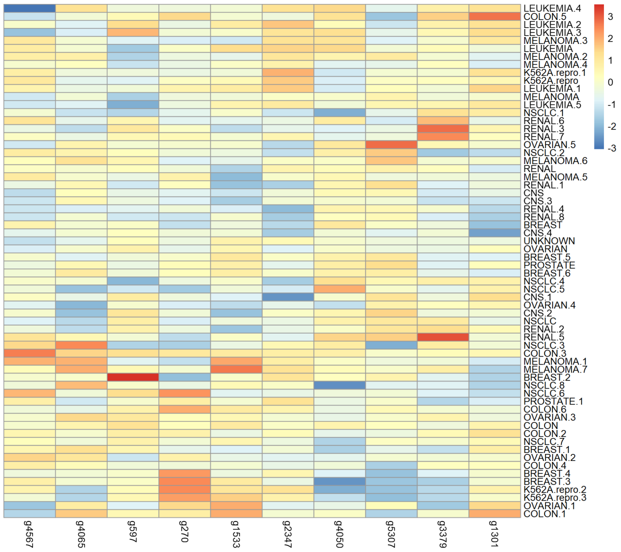
[3] C. Mukherjee and J. Zhang, "Compressibility: Power of PCA in Clustering Problems Beyond Dimensionality Reduction," *ResearchGate*. 2022. Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/360186068_Compressibility_Power_of_PCA_in_Clustering_Problems_Beyond_Dimensionality_Reduction

[4] C. Ding and X. He, "K -means clustering via principal component analysis," in *Twenty-first international conference on Machine learning - ICML '04*, Banff, Alberta, Canada: ACM Press, 2004, p. 29. doi: 10.1145/1015330.1015408.

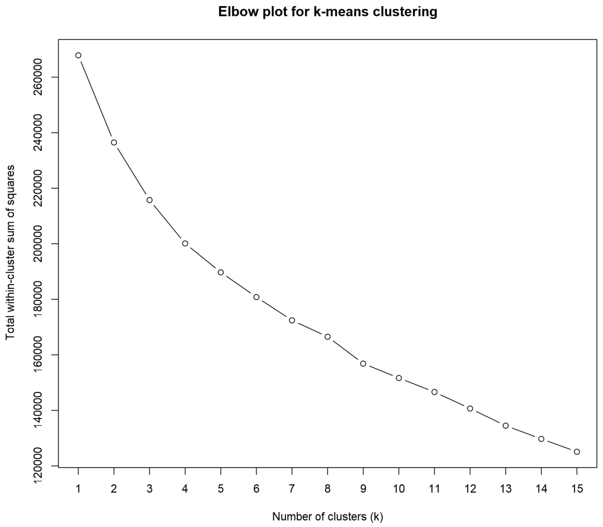
[5] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, Dec. 1971, doi: 10.1080/01621459.1971.10482356.

[6] W.-C. Chang, "On Using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 32, no. 3, pp. 267–275, 1983, doi: 10.2307/2347949.

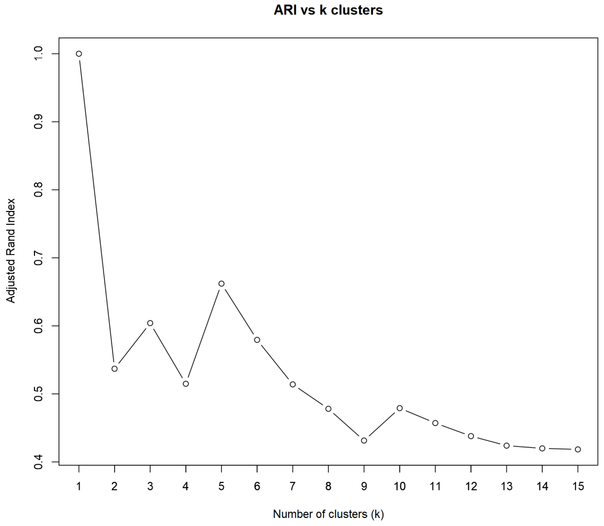
8 Appendix



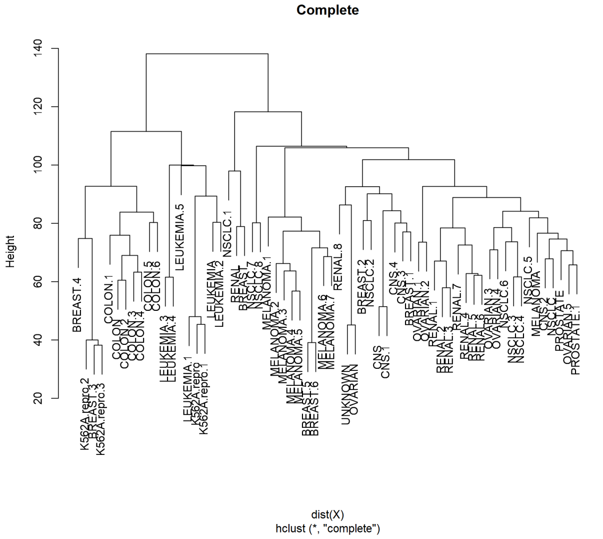
LEUKEMIA.4
COLON.5
LEUKEMIA.2
LEUKEMIA.3
MELANOMA.3
LEUKEMIA
MELANOMA.2
MELANOMA.4
K562A.repro.1
K562A.repro
LEUKEMIA.1
MELANOMA
LEUKEMIA.5
NSCLC.1
RENAL.6
RENAL.3
RENAL.7
OVARIAN.5
NSCLC.2
MELANOMA.6
RENAL
MELANOMA.5
RENAL.1
CNS
CNS.3
RENAL.4
RENAL.8
BREAST
CNS.4
UNKNOWN
OVARIAN
BREAST.5
PROSTATE
BREAST.6
NSCLC.4
NSCLC.5
CNS.1
OVARIAN.4
CNS.2
NSCLC
RENAL.2
RENAL.5
NSCLC.3
COLON.3
MELANOMA.1
MELANOMA.7
BREAST.2
NSCLC.8
NSCLC.6
PROSTATE.1
COLON.6
OVARIAN.3
COLON
COLON.2
NSCLC.7
BREAST.1
OVARIAN.2
COLON.4
BREAST.4
BREAST.3
K562A.repro.2
K562A.repro.3
OVARIAN.1
COLON.1
g4567
g4065
g597
g270
g1533
g2347
g4050
g5307
g3379
g1301
3
2
1
0
-1
-2
-3



Elbow plot for k-means clustering
Total within-cluster sum of squares
120000
140000
160000
180000
200000
220000
240000
260000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of clusters (k)

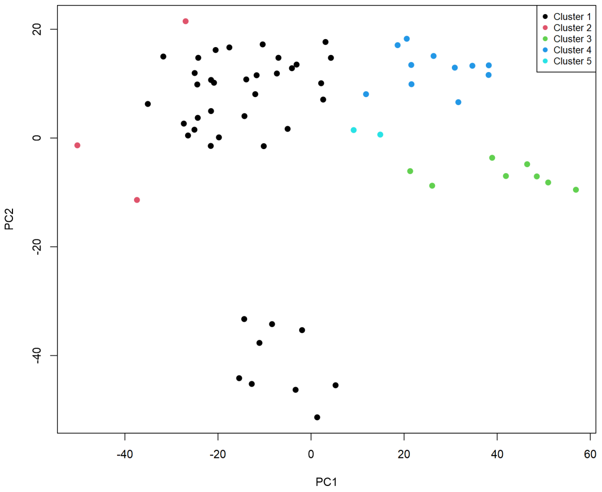


ARI vs k clusters
Adjusted Rand Index
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of clusters (k)

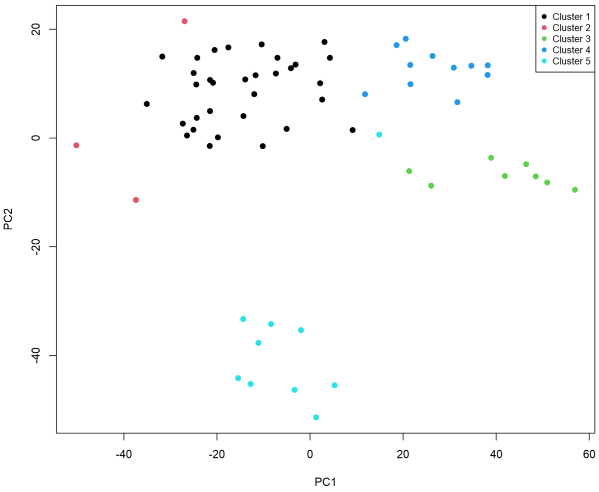


Complete
Height
20
40
60
80
100
120
140
BREAST.4
K562A.repro.2
BREAST.3
K562A.repro.3
COLON.1
COLON
COLON.2
COLON.3
COLON.4
COLON.5
COLON.6
LEUKEMIA.3
LEUKEMIA.4
LEUKEMIA.5
LEUKEMIA.1
K562A.repro
K562A.repro.1
LEUKEMIA
LEUKEMIA.2
NSCLC.1
RENAL
BREAST
NSCLC.7
NSCLC.8
MELANOMA.1
MELANOMA.2
MELANOMA.3
MELANOMA.4
MELANOMA.5
BREAST.5
BREAST.6
MELANOMA.6
MELANOMA.7
RENAL.8
UNKNOWN
OVARIAN
BREAST.2
NSCLC.2
CNS
CNS.1
CNS.4
CNS.3
BREAST.1
OVARIAN.1
OVARIAN.2
RENAL.1
RENAL.2
RENAL.3
RENAL.7
RENAL.4
RENAL.5
RENAL.6
OVARIAN.3
OVARIAN.4
NSCLC.6
NSCLC.3
NSCLC.4
NSCLC.5
MELANOMA
CNS.2
NSCLC
PROSTATE
OVARIAN.5
PROSTATE.1
dist(X)
hclust (*, "complete")

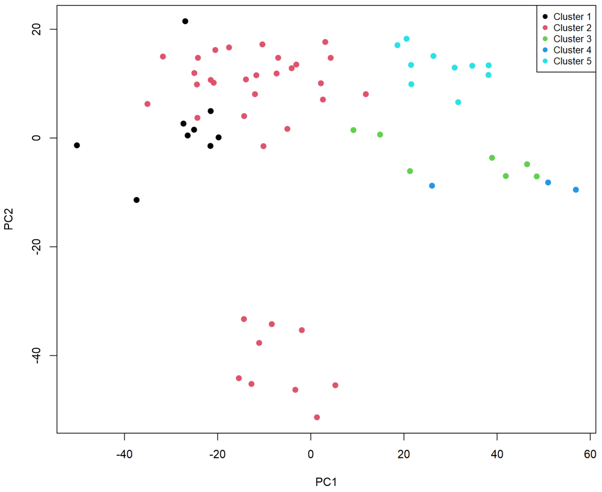
Complete linkage clusters (first 2 PCs, 5 clusters cut-off)



Complete linkage clusters (first 2 PCs, 5 cluster cut-off) on Z



Complete linkage clusters (first 2 PCs, 5 cluster cut-off) on Z_sp


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
PC2
PC1