**BIOL33021/66021 Assessment 2 (2023-24)**  
10724949

**INTRODUCTION**

The advent of single-cell analysis has emphasised the heterogeneity and complexity of cellular populations (Amir et al., 2013). Projects like Human Cell Atlas, The Cancer Genome Atlas Program and COVID-19 immune profiling are unravelling cellular diversity and dynamics in health and disease, leveraging single-cell sequencing's vast data. To be able to manage the massive amount of data intrinsic to this approach, Principal Component Analysis (PCA), k-means and hierarchical clustering became routinely employed (Lun et al., 2017; Shekhar et al., 2016). PCA, transforming correlated variables into principal components, reduces large dataset dimensionality, enhancing interpretability with minimal information loss (Jolliffe and Cadima, 2016). K-means clustering, partitioning observations into clusters based on the nearest mean, is essential for identifying cellular subpopulations and phenotypes (Hartigan and Wong, 1979; MacQueen et al., 1967). Hierarchical clustering, another powerful tool, uncovers natural groupings in cellular populations without predefined cluster numbers (Johnson, 1967).

This report explores PCA and clustering in single-cell flow cytometry data analysis, particularly for a COVID-19 immune profiling dataset (Lee et al., 2020). Using PCA, it visualizes cellular heterogeneity and identifies a marker gene, while k-means and hierarchical clustering categorize cells into subpopulations, demonstrating these methods' efficacy in elucidating single-cell dataset complexities.

**METHODS**

The study utilised single-cell flow cytometry data from a COVID-19 immune profiling study (Lee et al., 2020), focusing on lymphocyte populations. Data preprocessing involved removing debris and dead cells, followed by normalizing fluorescence intensities across various channels. The dataset then underwent log transformation to stabilise variance. PCA was conducted using the scikit-learn Python 3 library. The flow cytometry data, originally high-dimensional, were reduced to the first two principal components. K-means clustering, using scikit-learn, was applied to the PCA-reduced data to identify distinct cell populations. The number of clusters was determined using the Elbow method. Hierarchical clustering with Ward's method was then employed to further categorize cells into phenotypically distinct subpopulations.

**RESULTS**

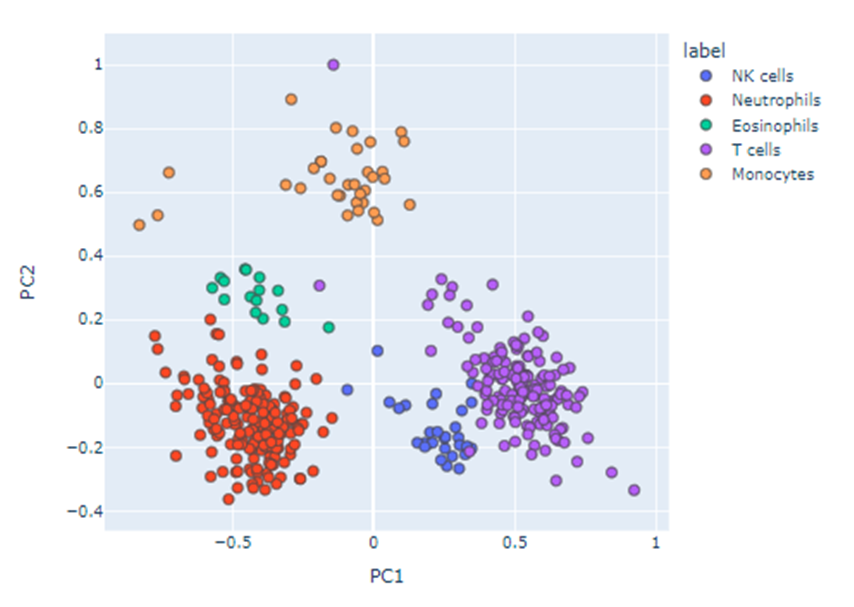
1. **PCA Visualization Analysis:**

The PCA analysis revealed a substantial variance in the dataset accounted for by the first two principal components, with the first component explaining 40% and the second 11% (Table 1). This significant reduction in dimensionality enabled effective visualization of complex cellular data.

**Table 1: Principal Component Variance Contributions in COVID-19 Cell Data Analysis** This table summarizes the variance accounted for by each principal component in the PCA analysis. The components are sequentially listed based on their contribution to explaining the total variance in the dataset. The table provides a quantitative measure of the significance of each principal component, crucial for understanding the dimensions that capture the most variability in the data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PC | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| % variance | 40 | 11 | 8 | 6 | 6 | 5 | 5 | 4 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

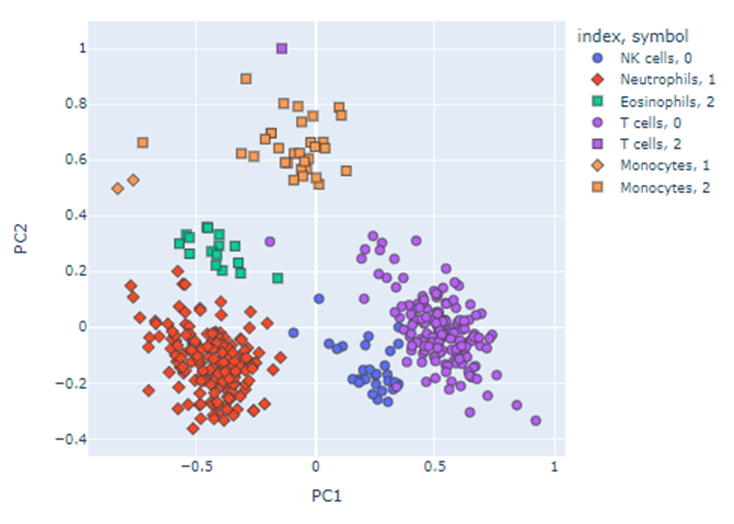
The 2D PCA visualization (Figure 1) showed varying degrees of separation among cell types. Monocytes formed a distinct cluster from others like lymphocytes and dendritic cells, highlighting PCA's effectiveness in capturing the variance that distinguishes monocytes. Other cell types, distinguishable, showed overlap, suggesting similarities in phenotypic expression. CD14 had significant positive loading on the second principal component, indicating its contribution to variance and potential as a monocyte marker gene.



**Figure 1: Two-Dimensional PCA Visualization of Cell Types in COVID-19 Immune Profiling** Legend: Figure 1 presents the PCA-based visualization of flow cytometry data, depicting the segregation of cell types in two-dimensional space. The x-axis and y-axis correspond to the first and second principal components, respectively, each contributing significantly to the variance in the dataset. Different cell types, identified based on specific markers, are represented by distinct colours. The legend in the corner specifies the percentage of total variance explained by each principal component. This visualization aids in identifying overarching patterns in the dataset and suggests the presence of distinct cell populations.

1. **K-Means Clustering Identification and Performance Assessment:**

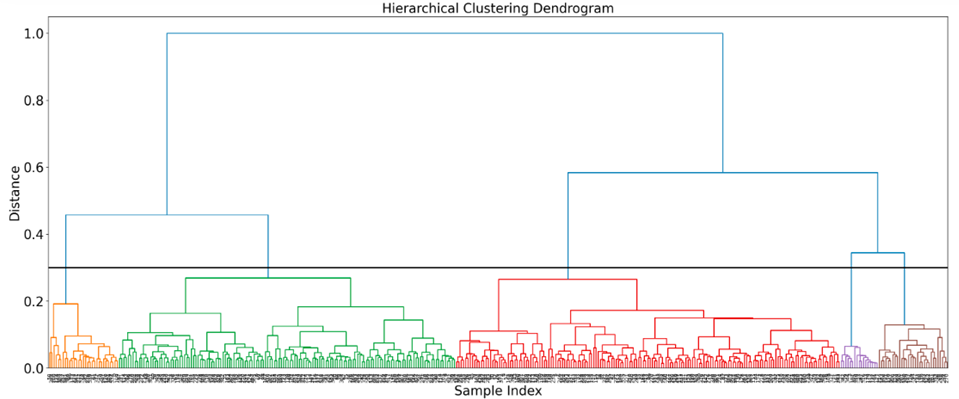
Applying k-means clustering to the dataset partitioned cells into groups based on flow cytometry data (Figure 2). This process effectively segregated cells into distinct clusters. The role of the k-means algorithm was to categorize these cells into clusters that represent potentially distinct cell types or states. Each cluster comprises cells with similar characteristics, thereby indicating that algorithm effectively segregated the cells into distinct groups, reflecting their phenotypic heterogeneity.  
  
The clustering performance was assessed using ARI, with a value of 0.849, indicating, indicating a high degree of similarity between the algorithm's clustering and the known cell type labels. This high ARI value suggests that the k-means clustering algorithm was successful in accurately identifying and grouping the different cell types, validating its efficacy in analysing single-cell flow cytometry data.



**Figure 2: K-Means Clustering Results and Ground Truth Comparison in COVID-19 Immune Cell Data** The graph illustrates the results of k-means clustering applied to COVID-19 immune cell data, presented in a PCA-reduced two-dimensional space. Each cell is assigned to one of the clusters, indicated by various colors, based on their expression profiles. Circles represent the centroids of the clusters identified by the k-means algorithm, highlighting the central tendencies of these groups. Additionally, squares are used to denote the ground truth labels, providing a visual comparison between the algorithmically determined clusters and the known cell types. The proximity and overlap between the circles and squares offer insights into the accuracy and efficacy of the k-means clustering in capturing the true biological variation of the cell populations. The clear delineation of clusters and their correlation with the ground truth labels underscore the algorithm’s capability in identifying distinct cellular phenotypes within the context of COVID-19 immune profiling. Notably, squares representing ground truth labels are present for certain cell types but absent for most T cells and NK cells. This absence might indicate that these cell types were not included in the ground truth labels of the dataset, thus their clusters are identified solely based on the k-means algorithm without external validation.

1. **Dependence of Clustering Performance on Linkage Method:**

Hierarchical clustering was performed using various linkage methods to explore how clustering performance depends on these methods. The dendrogram (Figure 3) was used for this assessment. The Ward method showed the highest ARI value of 0.977, indicating excellent performance in accurately clustering cells into biologically relevant groups. In contrast, other methods such as single, complete, average, weighted, and centroid linkage demonstrated lower ARI values (Table 2). These results suggest that the choice of linkage method significantly impacts the clustering performance, with the Ward method being the most effective in this particular dataset. The success of the Ward method, which minimizes total within-cluster variance, was crucial in generating cohesive and well-separated clusters, leading to a more accurate representation of the biological variance present in the single-cell data.



**Figure 3: Hierarchical Clustering Dendrogram for Cell Data Classification**   
This dendrogram illustrates the results of hierarchical clustering, showcasing the agglomerative process of grouping similar cell types based on their expression profiles. The vertical axis represents the distance metric, quantifying the dissimilarity between clusters, and each horizontal line denotes the fusion of two clusters. The length of the vertical lines reflects the degree of similarity between merged clusters. The dendrogram enables the identification of closely related cell groups and the understanding of the hierarchical structure within the cell population. In particular, it serves as a tool for identifying potential subgroups within well-known cell types, such as T cells and NK cells, based on their phenotypic similarities and differences. The branching patterns and the height at which branches merge give clues about the relative homogeneity or heterogeneity of these cell populations. Such insights are valuable for understanding the complexity of immune responses, particularly in the context of COVID-19, and can guide further biological investigation and validation.

**Table 2: Evaluation of Clustering Efficacy Using Adjusted Rand Index (ARI)** The table quantifies the effectiveness of various clustering methodologies using the ARI. ARI is a measure of the similarity between the clustering results and a set of predefined ground truth labels. A higher ARI value indicates better alignment with the known cell types, thereby validating the clustering approach. The table compares ARI scores across different clustering methods, offering a metric for their performance in distinguishing biologically relevant cell groups.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Linkage method | ward | single | complete | average | weighted | centroid |
| ARI | 0.9773163698735625 | 0.003425680525959406 | 0.08483702379655624 | 0.05218614649407378 | 0.05450860808677429 | 0.05450860808677429 |

**DISCUSSION**

In this study, we have successfully demonstrated the utility of principal component analysis (PCA) and clustering techniques in elucidating the intricate cellular dynamics present in COVID-19 immune profiling, highlighting both their strengths and the potential for further methodological enhancements to deepen our understanding of complex biological systems.

The PCA effectively reduced the complexity of high-dimensional flow cytometry data, facilitating the identification of distinct cell populations. The visualization through PCA revealed clear segregation among various cell types, such as monocytes, lymphocytes, and dendritic cells. The identification of CD14 as a potential marker gene for monocytes based on its significant positive loading on the principal components is a particularly notable outcome, aligning with known biological markers.

The k-means clustering algorithm, validated against ground truth labels, showed a high degree of accuracy, as indicated by ARI. This high level of concordance highlights the algorithm's utility in automated cell-type identification, which is vital in large-scale immunological studies like those associated with COVID-19.

The comparison of various linkage methods in hierarchical clustering provided insights into how different approaches affect clustering performance. The superior performance of the Ward method, in particular, underscores its suitability for datasets where the goal is to minimize within-cluster variance, a common requirement in flow cytometry data analysis.  
  
However, PCA's linear approach may not fully capture non-linear biological data relationships, potentially overlooking subtle but significant variations. This study's focus on descriptive statistics and clustering, without inferential statistical analysis, limits the comprehensiveness of the findings. The reliance on accurate ground truth labels for assessing clustering algorithms also poses a limitation.

Evaluating clustering methods in single-cell analysis relies on internal validation metrics and stability-based methods without ground truth labels. The silhouette score assesses cluster similarity, useful for determining optimal cluster numbers (Rousseeuw, 1987). Stability-based methods examine clustering consistency across dataset subsamples, indicating genuine data patterns (Ben-Hur et al., 2002). Consensus clustering, merging multiple results, reflects the data's biological structure (Monti et al., 2003).

Future work could integrate non-linear dimensionality reduction techniques, like t-SNE or UMAP, for deeper biological insights (van der Maaten and Hinton, 2008; McInnes et al., 2018). Applying consensus clustering could overcome single clustering approach limitations (Monti et al., 2003). Exploring unsupervised machine learning algorithms may uncover novel patterns in data lacking ground truth labels. In summary, the study effectively used clustering and visualization for COVID-19 cellular dynamics, but advanced computational methods could further enhance understanding of biological complexity.

**BIBLIOGRAPHY**

1. Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D, 2013. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol*, 31(6), pp. 545-552.
2. Shekhar K, Brodin P, Davis MM, Chakraborty AK, 2016. Automatic Classification and Visualization of High-Dimensional Cytometry Data Using Deep Learning. *Proc Natl Acad Sci U S A*, 113(26), pp. E3809-E3817.
3. Jolliffe IT, Cadima J, 2016. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, 374(2065), 20150202.
4. Lun ATL, McCarthy DJ, Marioni JC, 2017. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res*, 5, 2122.
5. Hartigan JA, Wong MA, 1979. Algorithm AS 136: A k-means clustering algorithm. *J R Stat Soc Ser C Appl Stat*, 28(1), pp. 100-108.
6. MacQueen J et al., 1967. Some methods for classification and analysis of multivariate observations. In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), pp. 281-297.
7. Lee JS, Park S, Jeong HW, Ahn JY, Choi SJ, Lee H, Choi B, Nam SK, Sa M, Kwon JS, Jeong SJ, Lee HK, Park SH, Park SH, Choi JY, Kim SH, Jung I, Shin EC, 2020. Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol*, 5(49), eabd6197.
8. Hinton GE, Salakhutdinov RR, 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786), pp. 504-507.
9. Rousseeuw PJ, 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, pp. 53-65.
10. Ben-Hur A, Elisseeff A, Guyon I, 2002. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, 7, pp. 6-17.
11. van der Maaten L, Hinton G, 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov), pp. 2579-2605.
12. McInnes L, Healy J, Melville J, 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
13. Monti S, Tamayo P, Mesirov J, Golub T, 2003. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2), pp. 91-118.