Introduction

Single-cell RNA sequencing (scRNA-seq) has become a transformative tool in understanding cellular heterogeneity. By measuring gene expression at the level of individual cells, scRNA-seq provides insights that bulk RNA-seq is incapable of, such as identification of rare or distinct cell types within a population. This capacity to capture transcriptional profiles at the single-cell level is crucial for understanding complex biological processes, including immune responses, cell differentiation, and disease mechanisms (Wolf et al., 2018). This study leverages machine learning (ML) approaches to annotate cell types within a scRNA-seq dataset from Peripheral Blood Mononuclear Cells (PBMCs).

PBMCs represents a diverse group of blood cells that include essential immune system components such as T cells, B cell, natural killer (NK) cells, and monocytes. These cells perform a range of vital immune functions, from recognising and eliminating pathogens to triggering adaptive immune responses. The dataset used in this analysis comprises of 68,000 cells, containing 4 different cell-types, providing a robust sample for developing and testing ML models (Abdelaal et al., 2019). The aim is to accurately label the cell types using computational approaches, addressing the limitations of manual annotations.

Traditionally, manual cell-type annotation in scRNA-seq relies on the expression of marker genes. For example, CD3 is a recognised marker for T cells, while CD19 is associated with B cells. Although effective for smaller datasets, this method becomes increasingly impractical for larger datasets and is susceptible to errors and biases, particularly when dealing with rare or poorly defined cell populations. To overcome these challenges, automated ML methods have become indispensable in scRNA-seq analysis (Ma et al., 2021), enabling researchers to scale up cell-type annotation while simultaneously improving the accuracy and consistency.

ML methods automate the process of identifying cell types by detecting patterns in gene expression data. These models use labelled data – cells with known identities – to predict cell types in further data. ML approaches effectively enhance consistency, reduce manual labour and offer scalability for larger datasets. However, their effectiveness hinges on factors such feature selection, preprocessing steps, and the choice of algorithms (Abdelaal et al., 2019). Proper preprocessing, including normalisation and filtering, is essential to ensure the data quality, as scRNA-seq data often contains a lot of noise.

In this report, two machine learning models – Support Vector Machine (SVM) and Neural Networks (NN) – are applied to a PBMC dataset to predict cell type annotations from gene expression data. A central focus of the study is a comparison of the performance of the two models. Evaluation metrics such as accuracy, precision and recall will be used to evaluate each model's effectiveness. SVM is suitable for high-dimensional data and handling complex classification tasks with the use of kernel functions. It is also robust to noise and is effective when there is a clear margin between classes, offering good interpretability of the class boundaries. On the contrary, NN particularly deep learning model, are better in capturing complex, non-linear relationships in the data. Given the scRNA-seq data often involves complex gene expression patterns, it is expected that the NN model will perform better than the SVM model in terms of the evaluation metrics. However, dimensionality reduction may be necessary to make NN training feasible, and this may lead to a loss of information that could negatively affect performance of the NN model. By comparing SVM and NNs, the study aims to identify the most effective approach for cell-type annotation in scRNA-seq data (Wolf et al., 2018). The findings will underscore the broader potential of machine learning in automating and refinement of scRNA-seq data analysis.

Data Set and Method

The dataset is a subset of the 68k PBMC scRNA-seq data from 10x Genomics, sequenced on the Illumina NextSeq 500 , and includes four cell types: CD8+ Cytotoxic T cells (immune response) (Domínguez Conde et al., 2022), Dendritic cells (antigen presentation) (Estipona, 2021), CD4+/CD45RA+/CD25- Naive T cells (adaptive immunity) (Pratt, 2023),

and CD14+ Monocytes (inflammation) (Estipona, 2020).The data is organised as an Annotated data object, with detailed gene expression and cell type annotations.

Data QC, Scaling, and Normalisation

This data was filtered in accordance with standard recommended procedures documented in the literature, which involved: removing cells with more than 5% reads from mitochondrial genes (Galow et al., 2021) or more than 50% reads from ribosomal genes (Grandi et al., 2022); removing cells with a number of genes outside the range of the median +/- 3x the mean absolute deviation (Lun et al., 2016); and, removing genes expressed in less than 3 cells because genes expressed in so few cells are not able to provide useful or meaningful information. Normalisation was performed across cells to normalise the reads per genes to give a constant library size of 10,000 reads per cell to avoid biases towards particular cells in the dataset (Hippen et al., 2021). Log-transformation was then applied with a pseudo count of +1 to reduce the difference in magnitude between highly and lowly expressed genes. Data was then scaled across genes to unit variance and a mean to zero to ensure genes are weighted comparably in downstream analysis, avoiding bias (Luecken & Theis, 2019).

Feature Selection

Feature selection was performed by selecting the genes with the greatest variance using ScanPy's *highly_variable_genes* method applied to the modified dataset (Satija et al., 2015; Stuart et al., 2019). This method identifies and annotates the genes that have much higher variance than is expected given their mean, so can be expected to separate out the genes in the dataset that have the greatest explanatory power for the relationships in the data and are therefore seen as most suitable for using as training data. This feature selection was validated by comparison with known biological markers (see Discussion).

Model training and testing

For both models, the dataset was divided into training and testing sets, with 80% of the data allocated for training and 20% for testing. Performance of both models was evaluated using precision, recall, and F1-score for each cell type.

Support Vector Machine (SVM)

In this study, a Support Vector Machine (SVM) model was implemented within a Conda environment to classify cell types. Initially, the SVM model was trained using a linear kernel to establish a baseline. Subsequently, a 5-fold cross-validation approach was employed, utilising GridSearchCV from the scikit-learn library, to optimise hyperparameters. A grid search was performed to identify the best combination of the regularisation parameter C and kernel type. The optimal hyperparameters were determined to be C = 10 and a radial basis function (RBF) kernel. Using these parameters, a new SVM model was trained.

Neural Network (NN)

A second model with a neural network architecture was trained on the data also for comparison. To reduce the time required to train the model to ensure it was practical to run on a laptop, dimensionality reduction was performed using PCA with scikit-learn's ARPACK solver to reduce the gene expression data from 2320 genes to 20 principal components. A feedforward NN was then trained on this reduced dataset, using MLPClassifier from scikit-learn, with an architecture consisting of a single hidden layer with 50 nodes. A ReLU activation and Adam solver was used, with a learning rate of 0.001. The model was trained for a maximum of 500 iterations, which was sufficient to see convergence of the loss function and thus reasonable model parameters. Cross-validation was also applied to fine-tune hyperparameters, ensuring effective performance.

Results

The initial quality assessment of the PBMC single-cell RNA-seq data showed a Pearson correlation coefficient of 0.972 between total reads per cells and the number of genes detected (Appendix - Figure 1), indicating consistent sequencing depth and uniform data quality across cells. This correlation suggests that the dataset is balanced and not skewed toward specific genes or subset of cells, which is crucial for reliable analyses.

Outlier cells were identified using the median absolute deviation (MAD) method, with filtering applied to remove cells with abnormally high or low gene counts. As shown in Figures 2 and 3, the distribution of detected genes per cell changed from a skewed shape before filtering to more closely resemble a Gaussian distribution after filtering, indicating the successful removal of low-quality cells. Gene-level filtering further reduced noise by excluding genes expressed in fewer than three cells, retaining features likely to capture meaningful biological variation.

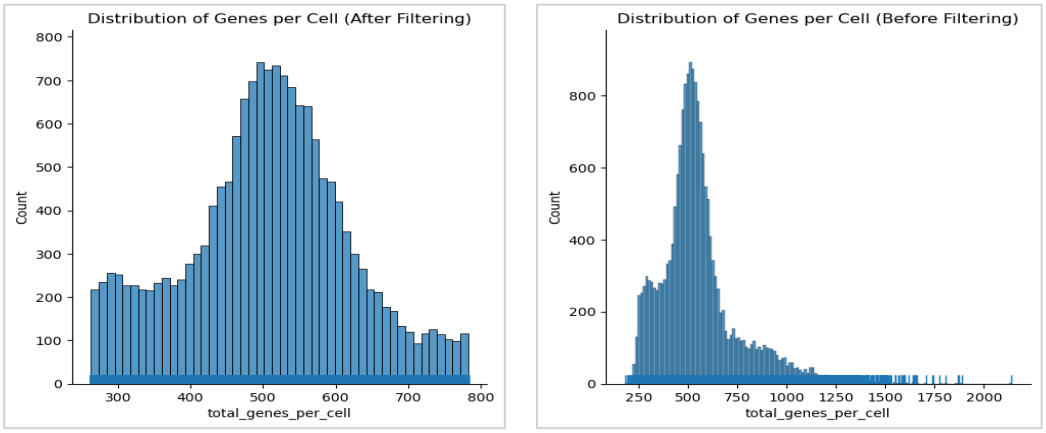

Figure 2 & 3: Median Absolute Deviation method (MAD) filtering applied to remove cells with low gene count.

Feature selection was based on identifying the most highly variable genes, which show greater variance in expression relative to their mean expression levels. Figure 4 in Appendix 2 illustrates the distinction between highly variable genes and less informative genes. This approach was motivated by studies such as the one by Le *et al.* (2022) which suggest that feature selection is a critical factor influencing the performance of ML models for single-cell RNA-seq analysis. By selecting highly variable genes, we focused on expression of genes likely to represent key biological signals, thereby improving the model's ability to accurately classify cell types.

Neural Network Classification - Machine Learning Cell Type Annotation

The neural network model achieved an accuracy of 92% in classifying the cell types (Appendix 3 - Table 1). The classification metrics indicated strong performance for CD8+ Cytotoxic T cells, which had an F1-score of 0.97, suggesting this cell type was well-recognised. This high performance can be attributed to the strong representation of CD8+ Cytotoxic T cells in the dataset, with 3,198 instances comprising approximately 65% of the total cells. In contrast, Dendritic cells exhibited lower performance with an F1-score of 0.63, likely due to their lower representation and overlapping expression profiles with other cell types (Estipona, 2021). with only 305 instances making up around 6% of the total dataset. The smaller number of examples for Dendritic cells increased the likelihood of misclassification, particularly given their overlapping expression profiles with other immune cell types.

The confusion matrix (Figure 6) further showed that most misclassifications occurred for Dendritic cells, which were frequently confused with other immune cell types. These results suggest that while the neural network effectively identifies major cell types, it struggles with rarer populations, highlighting the need for further optimisation of this model.

Support Vector Machine (SVM) Classification - Machine Learning Cell Type Annotation

The initial SVM model showed a high accuracy of 94%, with strong classification metrics for most cell types (Appendix - Table 2). Notably, the CD8+ Cytotoxic T cells were classified with near-perfect precision and recall, while Dendritic cells had lower recall, indicating challenges in identifying this cell type accurately.

Hyperparameter tuning, using grid search with 5-fold cross-validation, optimised the SVM's parameters (C=10, kernel='rbf'), leading to an improved accuracy of 95% (Appendix - Table 3). The best-tuned model showed increased recall for CD14+ Monocytes (0.98) and precision for Dendritic cells (0.94). However, the recall for Dendritic cells decreased to 0.55, indicating a trade-off in sensitivity after tuning. These findings align with the importance of feature selection and parameter optimisation highlighted by Le et al. (2022), which suggest that careful tuning can significantly impact model performance.

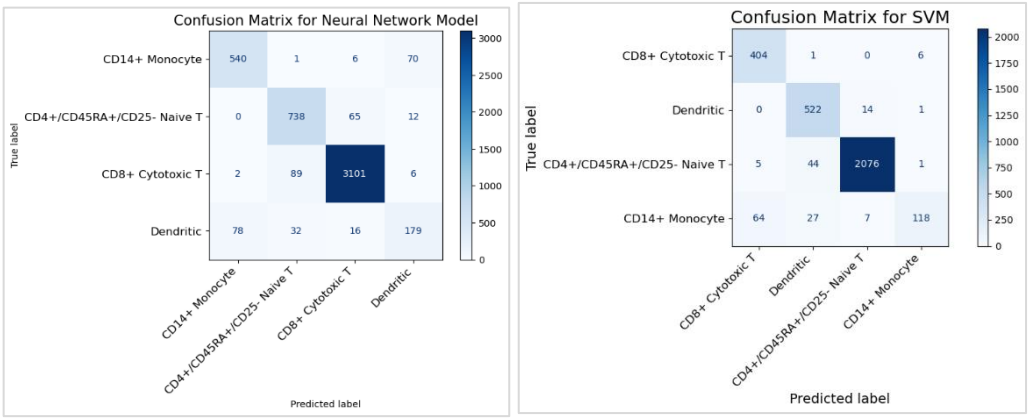

Figure 6 & 7: Neural Network and Support Vector Machine (SVM) Classification Confusion Matrix.

Discussion

The marker genes used for cell-type annotation closely align with established markers in the biological literature, which adds confidence to the cell-type assignments made in this analysis. For example, CD8+ Cytotoxic T cells were identified using markers such as CD8A, GZMB, and CX3CR1, all of which are well-documented indicators of cytotoxic activity (Xu et al., 2023). Similarly, markers used for Dendritic cells (DCs), including subtype-specific markers like CLEC9A and SIRPA (Estipona, 2021), match known subsets of DCs such as cDC1 and cDC2. The consistency of these markers with the literature suggests that the identified cell types in this dataset accurately reflect biologically relevant populations.

However, the lower performance for classification of DCs, which have overlapping expression profiles with other immune cells (Estipona, 2021), particularly in terms of recall, highlights the inherent challenges in accurately identifying this cell type. DCs are a diverse group, comprising multiple subtypes such as plasmacytoid DCs (pDC), conventional DCs (cDC1 and cDC2), and monocyte-derived DCs (mo-DCs), each with distinct marker profiles (Estipona, 2021). This heterogeneity can lead to overlaps in gene expression profiles with other immune cells, making it difficult for machine learning models to distinguish them from other cell types. The findings suggest that while the markers used were suitable for general DC identification, additional markers or multi-modal data might be needed for a more accurate classification of DC subtypes.

Both the NN and SVM models demonstrated strong classification accuracy, with SVM showing a slight edge after hyperparameter tuning. The literature indicates that careful feature selection significantly impacts model performance in single-cell RNA-seq analysis (Le *et al.*, 2022). In this study, selecting highly variable genes was used to focus on features with the most biological relevance, which is expected to improve classification outcomes in single-cell RNA-seq analysis.The choice of highly variable genes effectively reduced noise and enabled the models to concentrate on the key drivers of cell-type distinctions.

The neural network achieved an accuracy of 92%, with a particularly high F1-score of 0.97 for CD8+ Cytotoxic T cells, indicating that the model excelled in identifying these well-defined cell types. This suggests that the model is capable of effectively classifying cell types with clear marker expression patterns. However, the model struggled with rarer populations like Dendritic cells, where the F1-score (0.63) was notably lower. This limitation indicates that the model may have difficulty handling more complex expression profiles, likely due to the relatively lower expression of Dendritic cells in the dataset and their overlapping expression with other cell types (Estipona, 2021). These results suggest that additional data or model tuning may be needed to improve classification for less abundant cell types.

The initial SVM model had a strong performance with an accuracy of 94%. SVMs are well-known for their ability to handle high-dimensional data, making them suitable for single-cell RNA-seq analysis where each cell can be represented by thousands of gene expression features. After hyperparameter tuning (with parameters C=10 and kernel='rbf'), the model's accuracy improved to 95%, reflecting the importance of optimising model parameters to enhance performance. Similar to the Neural Network, the SVM showed excellent performance in classifying abundant cell types like CD8+ Cytotoxic T cells and CD14+ Monocytes. The precision and recall scores for these cell types were nearly perfect after tuning, showcasing the SVM's capability in learning from well-represented classes. The best-tuned model achieved higher precision for Dendritic cells (0.94), but with a trade-off in recall (0.55). This trade-off indicates that while the model became more confident in its predictions, it may have missed some Dendritic cells due to stricter decision boundaries imposed by the tuning process. This suggests that while SVM can be highly effective in identifying more common cell types, it also encounters challenges with rare cell types, particularly when optimising for overall accuracy.

In comparing the two models, SVM slightly outperformed the Neural Network in terms of overall accuracy, particularly after hyperparameter tuning. However, both models exhibited a common challenge in identifying Dendritic cells. The lower recall for Dendritic cells observed in both the neural network and SVM models underscores the difficulties in classifying this heterogeneous cell type. Dendritic cells share some gene expression patterns with other immune cells, such as Monocytes and T cells, making it challenging to draw clear boundaries between these populations. Additionally, the relatively lower representation of Dendritic cells in the dataset could have contributed to the models' reduced ability to detect them accurately. This highlights a limitation of both approaches, the need for better strategies to handle class imbalances in scRNA-seq datasets. While both models benefitted from feature selection based on highly variable genes, which helped capture biologically meaningful signals, additional methods such as data augmentation or class-specific weighting during training may further improve performance for rare cell types. A more balanced dataset can improve the classification performance of both SVM and Neural Networks, as handling class imbalance is crucial for accurately identifying less abundant cell types like Dendritic cells (LeCun et al., 2015)

Further exploration of model-specific tuning, such as incorporating class-weight adjustments or more advanced feature engineering techniques, may help address these classification challenges. For example, using ensemble methods that combine multiple classifiers or leveraging knowledge-based feature selection could improve recall for challenging cell types like Dendritic cells.

The results demonstrate that feature selection is a crucial step in single-cell RNA-seq analysis. By focusing on highly variable genes, the study was able to highlight genes that exhibit significant biological variability, which likely represent key functional differences between cell types. This approach aligns with previous research (Le *et al.*, 2022) that emphasises the impact of feature selection on machine learning model performance. In particular, using highly variable genes helped reduce noise, enabling the models to learn more discriminative patterns for cell-type classification. Overall, while both machine learning approaches showed strong classification capabilities, careful attention to feature selection and hyperparameter tuning was necessary to optimise their performance. The challenges encountered with Dendritic cell classification indicate that further refinement may be needed, potentially through leveraging more advanced modelling techniques tailored to rare cell populations. Future work can focus on optimising model architectures and exploring advanced techniques such as ensemble methods to address these issues and improve classification performance across all cell types.

References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*(1), 194. https://doi.org/10.1186/s13059-019-1795-z

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., … Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, *376*(6594), eabl5197. https://doi.org/10.1126/science.abl5197

Estipona, D. (2021, March 4). *A Guide to Dendritic Cell Markers*. Biocompare.com. https://www.biocompare.com/Editorial-Articles/572982-Dendritic-Cell-Markers/

Estipona, D. (2020, September 14). *A Guide to Monocyte Markers*. Biocompare.com. https://www.biocompare.com/Editorial-Articles/567890-A-Guide-to-Monocyte-Markers/

Galow, A. M., Kussauer, S., Wolfien, M., Brunner, R. M., Goldammer, T., David, R., & Hoeflich, A. (2021). Quality control in scRNA-Seq can discriminate pacemaker cells: the mtRNA bias. *Cellular and Molecular Life Sciences*, *78*(19-20), 6585-6592. https://doi.org/10.1007/s00018-021-03916-5

Grandi, F., Caroli, J., Romano, O., Marchionni, M., Forcato, M., & Bicciato, S. (2022). popsicleR: AR package for pre-processing and quality control analysis of single cell RNA-seq data. *Journal of Molecular Biology*, *434*(11), 167560. https://doi.org/10.1016/j.jmb.2022.167560

Hippen, A. A., Falco, M. M., Weber, L. M., Erkan, E. P., Zhang, K., Doherty, J. A., ... & Hicks, S. C. (2021). miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. *PLoS computational biology*, *17*(8), e1009290. https://doi.org/10.1371/journal.pcbi.1009290

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Le, H., Peng, B., Uy, J., Carrillo, D., Zhang, Y., Aevermann, B. D., & Scheuermann, R. H. (2022). Machine learning for cell type classification from single nucleus RNA sequencing data. *PLoS ONE*, *17*(9), e0275070. https://doi.org/10.1371/journal.pone.0275070

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

Lun, A. T., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, *5*. https://doi.org/10.12688/f1000research.9501.2

Ma, W., Su, K., & Wu, H. (2021). Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. *Genome Biology*, *22*(1), 264. https://doi.org/10.1186/s13059-021-02480-2

Pratt, C. (2023, June 16). *A Guide to Naïve T Cell Markers*. Biocompare.com. https://www.biocompare.com/Editorial-Articles/597618-A-Guide-to-Naive-T-Cell-Markers/

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, *33*(5), 495-502. https://doi.org/10.1038/nbt.3192

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., & Satija, R. (2019). Comprehensive integration of single-cell data. *cell*, *177*(7), 1888-1902. https://doi.org/10.1016/j.cell.2019.05.031

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15.

Xu, C., Prete, M., Webb, S., Jardine, L., Stewart, B. J., Hoo, R., He, P., Meyer, K. B., & Teichmann, S. A. (2023). Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell*, *186*(26), 5876-5891.e20. https://doi.org/10.1016/j.cell.2023.11.026

Appendix

Appendix 1: Table 1 - Neural Network Classification Report

	precision	recall	f1-score	support
CD14+ Monocyte	0.87	0.88	0.87	617
CD4+/CD45RA+/CD25- Naive T	0.86	0.91	0.88	815
CD8+ Cytotoxic T	0.97	0.97	0.97	3198
Dendritic	0.67	0.59	0.63	305

accuracy			0.92	4935
macro avg	0.84	0.84	0.84	4935
weighted avg	0.92	0.92	0.92	4935

Appendix 2: Table 2 – Support Vector Machine (SVM) Classification Report

	precision	recall	f1-score	support
CD14+ Monocyte	0.89	0.91	0.90	411
CD4+/CD45RA+/CD25- Naive T	0.87	0.91	0.89	537
CD8+ Cytotoxic T	0.98	0.98	0.98	2126
Dendritic	0.73	0.66	0.69	216

accuracy			0.94	3290
macro avg	0.87	0.86	0.87	3290
weighted avg	0.94	0.94	0.94	3290

Appendix 3: Table 3 - Support Vector Machine (SVM) Classification Report after hyperparameter tuning (using 5-fold cross-validation)

	precision	recall	f1-score	support
CD14+ Monocyte	0.85	0.98	0.91	411
CD4+/CD45RA+/CD25- Naive T	0.88	0.97	0.92	537
CD8+ Cytotoxic T	0.99	0.98	0.98	2126
Dendritic	0.94	0.55	0.69	216

accuracy			0.95	3290
macro avg	0.91	0.87	0.88	3290
weighted avg	0.95	0.95	0.95	3290

Appendix 4: Correlation between Gene Count and Read Counts per Cell

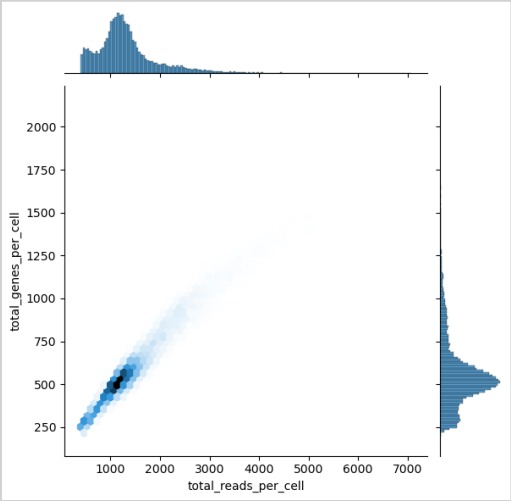

Figure 1: Quality control assessment indicating strong correlation between total reads per cell and the number of genes detected.

Appendix 5: Normalised and non-normalised dataset showing distinction between highly variable genes and less informative genes

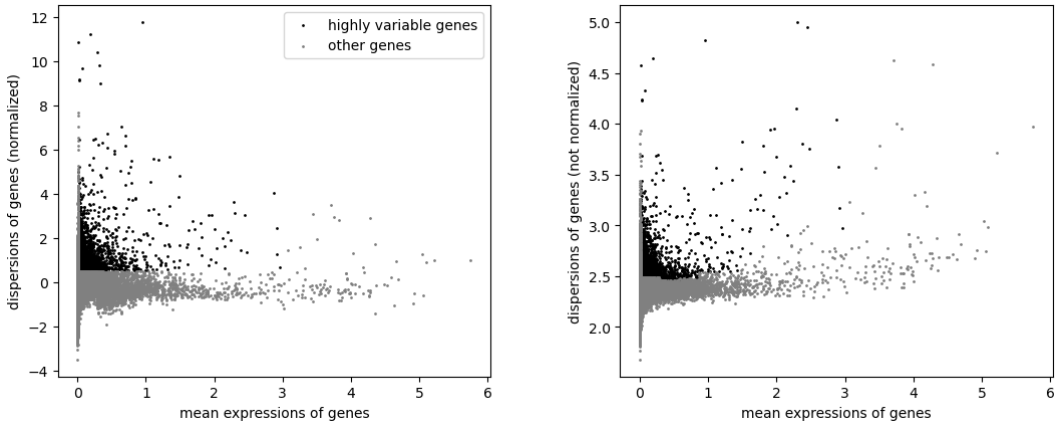

Figure 4 & 5: Visualisation showing distinction between highly variable genes and less informative genes.