Clustering of Single-Cell RNA Sequencing Data via Latent Space and Classification of Cancerous Cells with Marker Gene Identification and Class Imbalance Downsampling

2021 Summer STEM Institute

Minnie Liang

Abstract

Single cell RNA sequencing (scRNA-seq), a form of DNA sequencing technology, has emerged as a powerful tool to examine the gene expression of cells at an individual level and discover new cell subpopulations. With scRNA-seq technology, researches are now able to isolate and sequence the genetic material of single cells, opening the world of opportunities for cell to cell gene expression analysis compared to previous bulk RNA sequencing techniques. In this study, cell level analysis of scRNA-seq pancreatic cancerous tissue is performed via cell type clustering with scVI and compared to PCA dimension reduction. Cells are then classified as cancerous or noncancerous. Several machine learning methods were tested in addition to the affect of downsampling and feature selection methods. When downsampling the majority class, noncancerous cells, there was a 28.7% increase in f1 score compared to not downsampling. The effect of using the union or intersection of marker genes from cancer cluster A and B as features for the model was tested too, and a 28.7% f1 score increase was observed. In addition to cell level analysis, gene level analysis was performed by identifying the feature important genes for our model. Identified genes were then analyzed by cell type expression and validated through literature. In the process of combining cell level and gene level analysis, this study also aims to address the complexity of scRNA-seq and how to cope with the ultra high dimensionality and inflation of zero values.

1. Introduction

DNA sequencing has become a very popular technique for researchers, with many claiming it has the power to be equivalent to microscope impact in the future (Vrahatis et al., 2020). Single cell RNA sequencing (scRNA-seq), one of the forms of DNA sequencing technology, has emerged as a powerful tool (Vrahatis et al., 2020). scRNA-seq technology allows researchers to isolate and sequence the genetic material of single cells (Haque et al., 2017). This has opened the world of opportunities compared to previous bulk RNA sequencing, as now specific gene expression in specific cell types is able to be examined. The advancements made with single cell transcriptomes en-mass has allowed for identification of cell subpopulations and differential gene expression analysis (Vrahatis et al., 2020). It has already made a huge impact to a multitude of varying research fields, such as development (Semrau et al., 2017), autoimmunity (Gaublomme et al., 2015), and cancer (Patel et al., 2014).

scRNA-seq has a wide range of promising capabilities but also has a numerous number of computational challenges because of its large-scaled generated data. As sequencing costs are ever decreasing, the volume and complexity of the data generated by sequencing technologies continue to increase (Vrahatis et al., 2020). Because of this, major computational challenges have been posed at the cell level, especially dealing with the ultra-high dimensionality of scRNA-seq data. This study focuses on the classification, clustering, and visualization of cell subpopulations in scRNA-seq data of pancreatic ductal adenocarcinoma (PDAC) tissue. The analysis of gene expression and cell subpopulations of single cells in PDAC tissue opens the doorway to understanding the transcriptome in cells at an individual level. This can illuminate their role in cellular functions and how gene expression responds to beneficial or harmful states (Miao et al., 2018).

---

[*] *Research project conducted during the 2021 Summer STEM Institute.* The Summer STEM Institute (SSI) is a six-week virtual summer program where high school students learn how to design and conduct data science and other computational research projects. To view other SSI distinguished projects, please visit www.summersteminstitute.org.

2. Literature Review

2.1. Bulk RNA sequencing

Until very recently, the majority of biological experiments conducted used bulk cell sequencing, where a pool of cell populations is analyzed (Stegle et al., 2015). Many successful and well established studies have been done involving technologies like microarrays (Auer & Doerge, 2010) and RNA-seq (Gong et al., 2018). However, while these studies have been able to make a commendable significance in the field of biomedicine, bulk-cell measurements can often contain conflicting information between individual cells within these populations that is critical to know. Moreover, when obtaining information from a bulk pool of cells, unique biological properties of an individual cell such as its size, gene expression level, and biomarker expression are hidden. As seen in Figure 1, when looking at gene X, there is no change in expression when performing bulk RNA sequencing because it is obtaining information from a pool of cells. However, with scRNA-seq, the expression of gene X can be analyzed within different cell types and by clustering and read analysis, it can be concluded that expression is affected in cell type b only. Bulk cell sequencing studies are based on the assumption that cells from a given tissue are homogeneous, and thus, these studies are likely to miss out on important cell-to-cell variability that single-cell sequencing is able to provide.

2.2. Classification with scRNA-seq

Computationally, the main challenges of scRNA-seq data can be split into two areas, the cell level and the gene level analysis. In the former analysis, machine learning approaches like classification and clustering are used to assign cell classes and explain cellular heterogeneity. A promising use of scRNA-seq technology is classifying individual cells into pre-specified classes, whether it be cell type, cell-cycle stages, or cell quality. DropLasso, a method that deals with dropout noise and zero inflation of scRNA-seq data, proposed a new loss function based on stochastic gradient approach to classify breast cancer cell subtypes (Khalfaoui & Vert, 2018). Another method of interpreting single cell data is classifying cell cycles and phases. Buettner et al. used cell cycles to correlate with genes to identify the existence of new cell subpopulations in mouse embryonic stem cells, a rapidly cycling type of cell where a great fraction of its variability can be attributed to the cell cycle (Buettner et al., 2015). The authors in another study (Scialdone et al., 2015) developed an algorithm for the prediction of G1, S, or G2M phase using scRNA-seq data and a gene pairwise scoring scheme based on relative expression of gene pairs in a cell. Another way of analyzing scRNA-seq data is through classification of cell quality. In a study by Ilicic et al, the researchers were able to detect low quality cells that could not be identified visually by examining biological and technical features (Ilicic et al., 2016).

2.3. Clustering with scRNA-seq

scRNA-seq datasets are typically large and extremely high dimensional, with hundreds of thousands of genes and cells. A large dataset size ensures that analyses will have more power and capability of detecting hidden cell subpopulations. However, this high dimensionality data also imposes key limitations because of its size and complexity. Because of the low levels of RNA obtained from a single cell, scRNA-seq generally exhibits higher levels of noise and more zero counts than sequencing data from bulk-cell populations (Andreu-Perez et al., 2015). This zero inflation seen in scRNA-seq datasets introduce a numerous number of computational challenges. There are several statistical methods available for imputing zeros (Van Dijk et al., 2018), but they all rely on pre-existing cell to cell or gene to gene correlations to infer the imputed value. Recently, a handful of clustering approaches have been proposed for scRNA-seq data including clustering workflows, dimension reduction approaches, application of well established clustering tools, and tailored similarity measurements. The software packages Seurat (Macosko et al., 2015), scanpy (Wolf et al., 2018), and SINCERA (Guo et al., 2015) have implemented a full clustering workflow for scRNA-seq data including preprocessing steps like filtering, normalization, and quality control.

2.4. Gene Level Analysis with scRNA-seq

The second pillar of analysis is based on gene expression and includes approaches such as differentially expressed genes and gene expression dynamics. Kharchenko et al used a noise-tolerant probabilistic model based off of scRNA-seq data to detect differential expression signatures of genes (Kharchenko et al., 2014). A multitude of analysis tools have also been developed for differential gene expression analysis of scRNA-seq including SCDE (Kharchenko et al., 2014), MAST (Finak et al., 2015), scDD (Korthauer et al., 2016), D3E (Delmans & Hemberg, 2016), Monocle2 (Qiu et al., 2017), SINCERA (noa), DEsingle (Miao et al., 2018), and SigEMD (Wang & Nabavi, 2018). In a study by Briggs et al, time series of single-cell transcriptome measurements was used to reveal dynamic features of genes in cell states of frog embryos. The study was able to reveal conserved and divergent features of vertebrate early developmental gene expression programs (Briggs et al., 2018). Another study developed CONICS (COpy-Number analysis In single-Cell RNA-Sequencing), a software tool created for mapping gene expression from scRNA-seq to tumor clones and phylogenies (Müller et al., 2018).

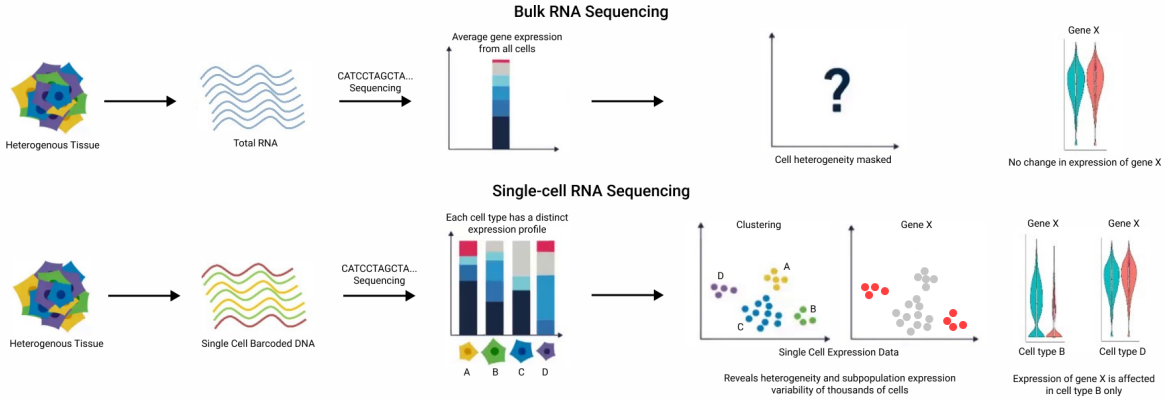

Figure 1. Diagram of the differences between bulk RNA sequencing and Single-cell RNA sequencing. scRNA-seq is able to examine gene expression of cells at an individual level, whereas cell heterogeneity and subpopulations are hidden for bulk RNA seq.

2.5. Purpose

As seen, a multitude of technologies have been made to analyze heterogeneity in cell populations. With the development of scRNA-seq, it has provided researchers a powerful capability to identify cell subpopulatons and gene activity at an individual level compared to prior bulk sequencing (Poirion et al., 2016). In this study, not only is the cell level of scRNA-seq pancreatic cancerous tissue analyzed via clustering and classification, but also at the gene level with differential expression and identification of marker genes. In the process of combining cell level and gene level analysis, this study also aims to address the ultra-high dimensionality and complexity of scRNA-seq and how to cope with this curse of dimensionality and zero inflation.

3. Methodology

3.1. Dataset

The scRNA-seq dataset used in this study is from a patient's primary pancreatic cancer tissue and was collected by the Institute for Computational Medicine at NYU Langone Health (Moncada et al., 2020). The dataset consists of 19738 genes and 20 cell types. It was collected to be combined with a spatial transcriptomics dataset and the overlap of the two datasets' differentially expressed genes was analyzed to uncover hidden cell types and spatial locations.

3.2. Workflow

First, the scRNA-seq dataset was clustered by the cell transcriptomes' latent space with scVI. The results from this was compared to PCA dimension reduction, a common technique performed on scRNA-seq to combat ultra-high dimensionality and inflation of zero values. Next, marker genes of the cancerous cell types was identified through differential expression analysis. Using these marker genes, a binary classification model was built to predict if a cell is cancerous or noncancerous. Downsampling the majority class and different subsets of the marker genes as model features were experimented. Then, the most important genes the models uses in its predictions were found and their cell type expression was analyzed.

3.3. Clustering

First, using pre-built functions in scanpy (Wolf et al., 2018), the top 1200 highly variable genes out of the total 19738 were selected by categorizing genes by mean and standard deviation expression of genes. Then, a latent representation of these 1200 genes was created using scVI-tools (Lopez et al., 2018). scVI learns a nonlinear embedding of the cells through a neural network that maps the latent variables to the parameters of the zero-inflated probability (ZINB) distribution. This transforms the data into a low-dimensional latent vector of normal random variables, which helps combat the high dimensionality of scRNA-seq data. This latent represenation of the transcriptome of each cell was then inputted into PCA, tSNE, and UMAP to cluster cell types. To compare the results of clustering via scVI latent space, cluster maps were made using PCA and PCA dimension reduction on top of tSNE and UMAP.

3.4. Classification

Due to few data instances of the more obscure cell types, classification of cell types was converted into binary classification, where 1 stood for a cancerous cell (either from the Cancer Clone A cell type or the Cancer Clone B cell type) and 0 stood for noncancerous and encapsulated the rest of the 18 cell types. There was a significant dataset imbalance, with 85% of the data set being noncancerous, leaving only 15% to be labelled 1, cancerous.

Then, through scVI-tools, marker genes were found by look-

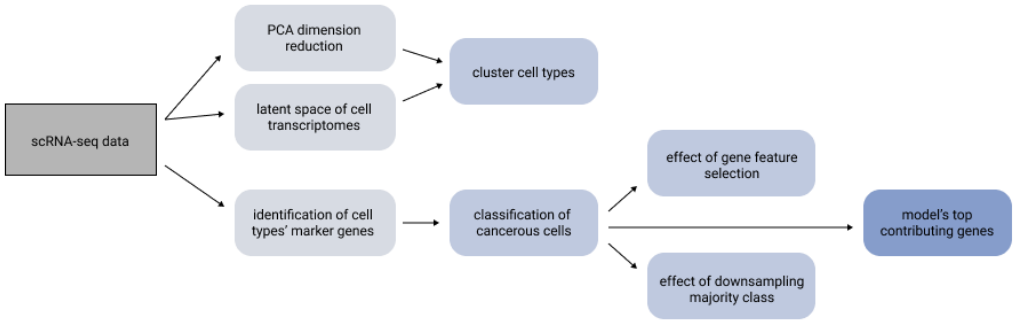

Figure 2. Layout of this study's work.

ing at differential expression and performing hypothesis testing based on latent representation. Genes with a mean log fold change of greater than 0, a bayes factor (1 vs. all for each cell type) of greater than 3, and a proportion of non-zero expression in cell type greater than 0.1 were determined as the marker genes.

Combined, there were a total of 156 marker genes from Cancer Clone A cell type vs Rest of cells and for Cancer Clone B cell type vs Rest of cells. As seen in Figure 3, set A and set B share an overlap of 34 genes.

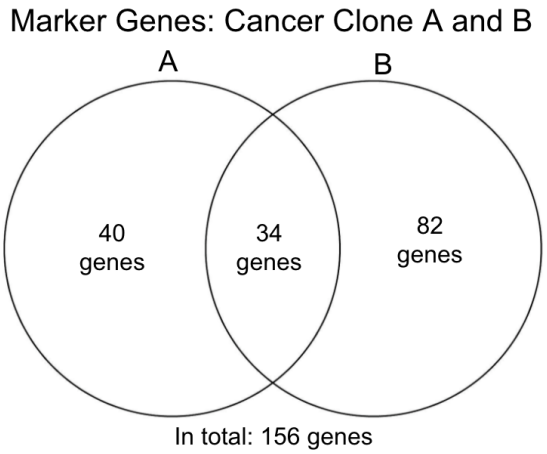

Figure 3. Venn diagram of the marker genes of set A (Cancer Clone A cell type vs Rest of cells) and set B (Cancer Clone B cell type vs Rest of cells)

The classification model was experimented on using A ∪ B or A ∩ B as model features. They were then dimension reduced with PCA and scaled to a range between 0 and 1 before being input into the model. The effect of downsampling or no change to the dataset was also studied. Four different models, Logistic Regression, Gradient Boosting, Random Forest, and Support Vector Machine (SVM) were implemented. Hyperparameter tuning was done by gridsearch with f1 scoring.

4. Clustering Results

4.1. PCA Dimension Reduction

After reducing the dataset into only the top 1200 highly variable genes, PCA, tSNE, and UMAP models on top of PCA dimension reduction were used to cluster cell types. As seen in Figure 4, PCA on tSNE and PCA on UMAP do much better resolving cell resolutions than just PCA. The cluster groups for PCA on UMAP are much less sparse compared to the clutering of PCA on tSNE. All four of the Ductal cell cluster groups are close to each other. PCA on tSNE works better than PCA on UMAP at resolving cell resolution between Ductal- terminal ductal like, Ducutal-MHC Class II, and Ductal-APOL1 high/hypoxic cell types. Cancer clone A and Cancer Clone B cell types also are clustered near each other.

4.2. scVI Latent Space

Instead of using PCA to transform the reduced dataset consisting of only the top 1200 highly variable genes, the latent space of these genes are calculated and inputted into PCA, tSNE, and UMAP models. As shown in Figure 5, the scVI on tSNE and scVI on UMAP do better at clustering than scVI on PCA. A similar trend was seen when doing dimension reduction with PCA. The clustering using scVI on PCA and scVI on tSNE closely resembles the corresponding PCA dimension reduction version. Similar to the trend seen in the PCA dimension reduction clustering, the cluster groups for scVI on UMAP are less sparse than the clustering for scVI on tSNE.

The most drastic change in clustering between using PCA dimension reduction versus scVI latent space is UMAP. PCA on UMAP had very dense clusters whereas with scVI on UMAP, the cell resolution between the different ductal cells improves greatly. The pink (Ductal- terminal ductal like), brown (Ducutal-MHC Class II), and red (Ductal-APOL1 high/hypoxic) clusters are now more distinguished and separated from one another than before. The comparisons

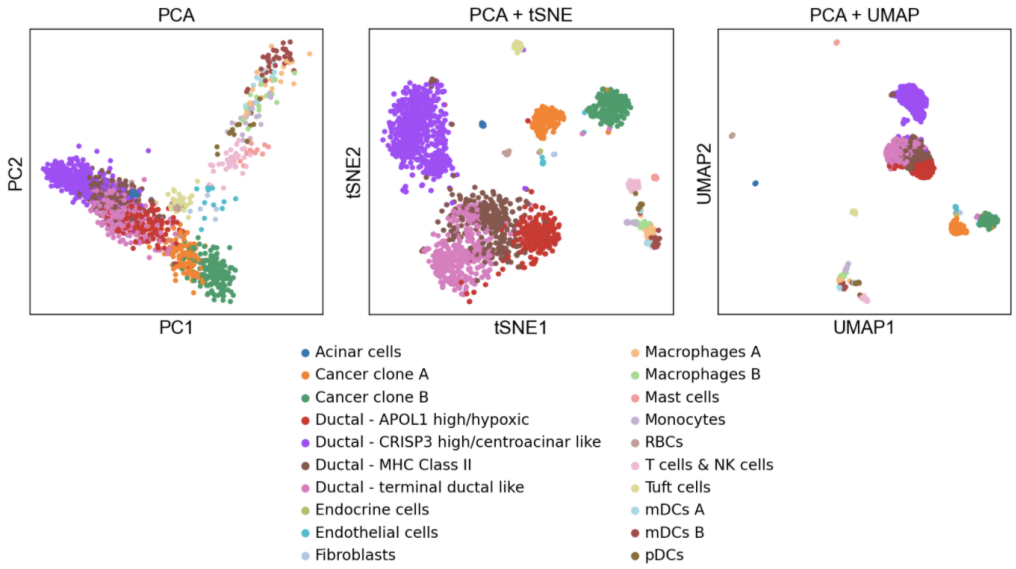

Figure 4. Clustering of cell types using PCA dimension reduction

between Figure 4 and Figure 5 seems to suggest that using the latent space to cluster cell types improves resolution compared to PCA dimension reduction.

5. Classification Results

Aside from clustering, classification using marker genes from cancer clusters as model features was used to predict if a cell is cancerous or not. The top contributing genes for our model were then identified and its cell type expression was analyzed.

5.1. Gene Feature Selection: A ∪ B or A ∩ B

To predict cancer cell types, the features of the model are from the marker genes of Cancer Clone A and the marker genes of Cancer Clone B. The effect of using the union of the set of marker genes from Cancer Clone A and B versus the effect of using the intersection of marker genes from Cancer Clone A and B was studied. A ∪ B resulted in 156 marker genes as features of the model whereas A ∩ B had 34 genes (Figure 3).

5.2. Data Label Imbalance: Downsampling Vs. No Change

Four different models were implemented: logistic regression, gradient boost, random forest, and support vector machine. The labels for the model were either 1 (15% of the dataset) or 0 (85% of the dataset), cancerous cell type or noncancerous cell type, respectively. The data was split into 80% for cross validation training and 20% for testing. Gridsearch 5-fold cross validation with f1 scoring was used to tune the hyperparameters of all four models using the training set. Two different factors were tested with these settings. One was downsampling the dataset by randomly removing data entries with 0 as the cell type label until the percentage of 0 labels was 59% and the percentage of 1 labels was 41%. The other case was leaving the dataset as is, with 85% of the labels being noncancerous cell type and 15% of the labels being cancerous cell type. While the cross validation dataset was experimented with downsampling, the test dataset was left unchanged, meaning that it still suffered from a big data imbalance, with 323 noncancerous cells and only 63 cancerous cells.

The effects of using A ∪ B or A ∩ B as the model's feature genes were tested alongside downsampling the majority class or not downsampling. So in total, four different experiments were run: Downsampling Vs. No Change on A ∩ B, Downsampling Vs. No Change on A ∪ B, Downsampling Vs. No Change on A ∪ B, and Downsampling Vs. No Change on A ∩ B.

5.3. Downsampling Vs. No Change on A ∩ B

With gridsearch 5-fold cross validation, the validation f1 scores and error bars were graphed in Figure 6 for all four models. The error bars do not overlap for gradient boost, random forest, and support vector machine. This suggests that downsampling the majority class on the union of marker genes from Cancer Clone A and B gives better scores than not downsampling and appears to be statistically significant. Using the best parameters found via gridsearch 5-fold cross validation, the entire 80% training dataset was fit onto the model and predicted on the 20% test dataset. The bar graph

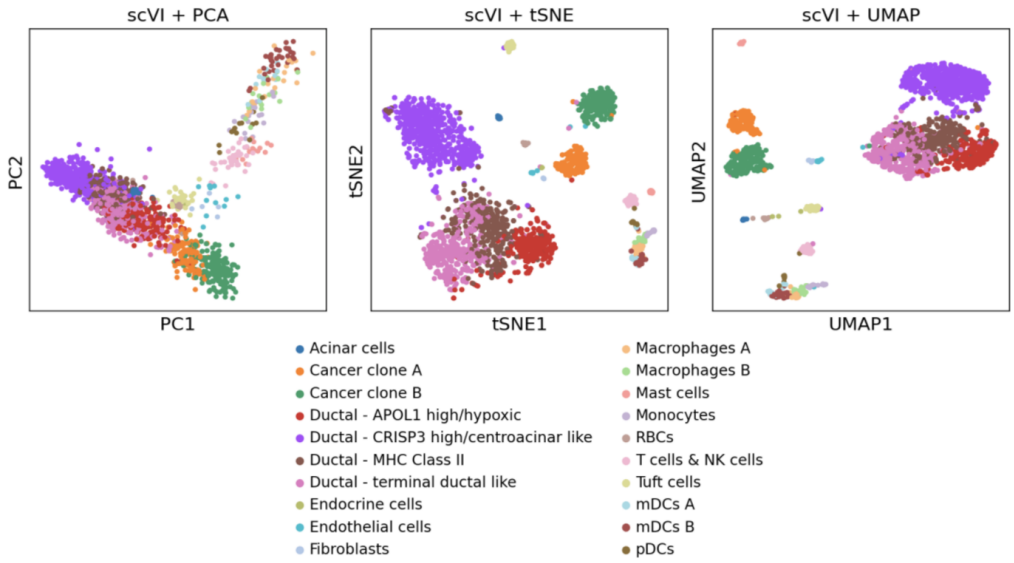

Figure 5. Clustering of cell types using scVI latent space

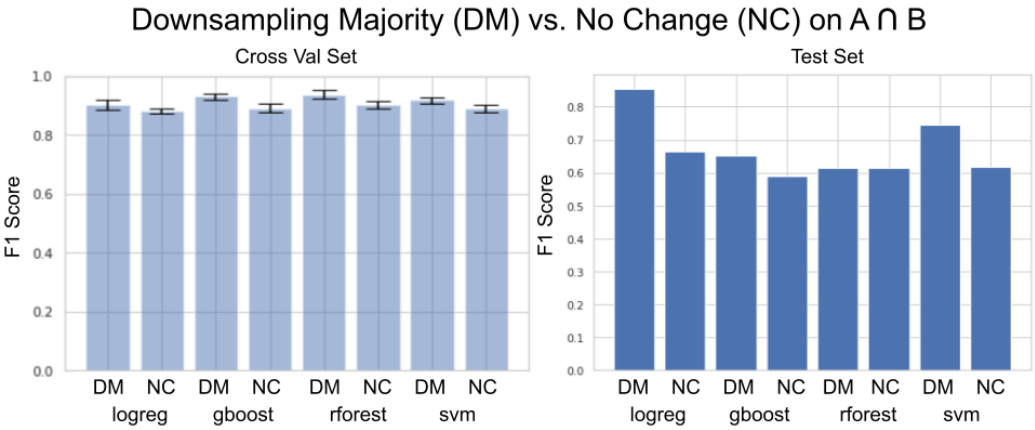

Figure 6. Bar graph of the f1 scores retrieved from four different machine learning models when downsampling the majority class versus not downsampling. The models' features is the intersection of marker genes from cancer clone A and B (A ∩ B).

of the test set f1 scores in Figure 6 show that downsampling the majority class does better than not downsampling, for all four of the models.

Based on the length of the error bars obtained from grid-search cv (Figure 6), the downsampling on logisitic regression model and downsampling on random forest seem to have the most variability in f1 scores. While downsampling on random forest seems to perform the best on the cross validation set, the best test score was gotten through downsampling the dataset with a logistic regression model. Here, a f1 score of 0.85 was achieved whereas the downsampling on random forest only achieved a score of 0.61.

As seen in Figure 7, the logistic regression model is able to predict quite well on the test set, despite the big class imbalance of cancerous cells the test set has. The next best test score achieved was using downsampling on support vector machine with a f1 score of 0.75. The SVM model has much less precision in predicting cells as cancerous (Figure 8).

5.4. Downsampling Vs. No Change on A ∪ B

The same process from section 5.3 was applied, except the set of features inputted in the four different type of models was the union (instead of intersection) of marker genes from Cancer Clone A Cluster and Cancer Clone B Cluster.

The error bars do not overlap for any of the four models,

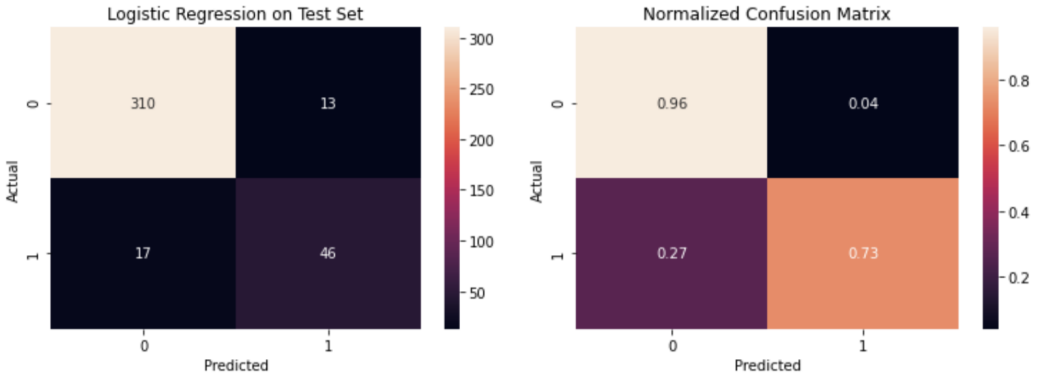

Figure 7. Confusion matrix and normalized confusion matrix of test set predictions using A ∩ B, downsampling on logistic regression.

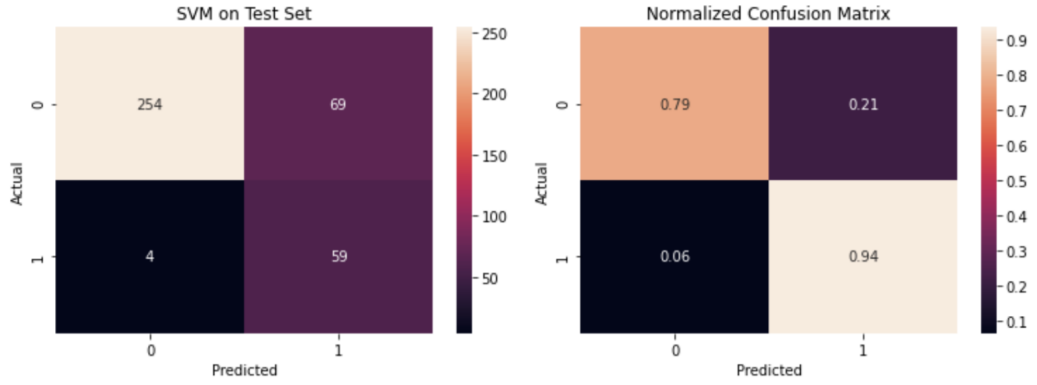

Figure 8. Confusion matrix and normalized confusion matrix of test set predictions using A ∩ B, downsampling on support vector machine.

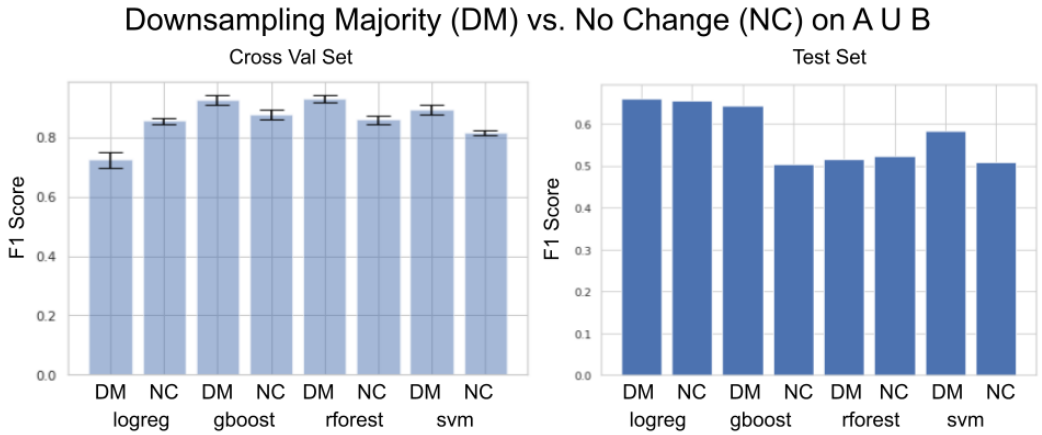

Figure 9. Bar graph of the f1 scores retrieved from four different machine learning models when downsampling the majority class versus not downsampling. The models' features is the union of marker genes from cancer clone A and B (A ∪ B).

suggesting that downsampling the majority class scores better than not downsampling for gradient boost, random forest, and SVM. Interestingly enough, for the logistic regression models, the nondownsampling dataset scored better than the downsampled class for the cross validation set.

The model, with parameters found via gridsearch 5-fold cv, was then predicted on the test dataset. Downsampling the number of noncancerous cell instances of the dataset performed better for 3 out of the 4 models: logistic regression, gradient boost, and SVM. An interesting note is that

while downsampling on logistic regression performed worse than not downsampling on logistic regression for the cross validation set, the reverse happened for the test set. The top performing models on the cross validation set were downsampling on gradient boost and downsampling on random forest, with f1 scores of 0.925 and 0.929, respectively. For the test set, the best scoring models were downsampling on logistic regression, no change on logistic regression, and downsampling on gradient boost, with f1 scores of 0.66, 0.65, and 0.64, respectively. The logistic regression had the highest variability based on length of the error bars in Figure 9, which may be the reason why this model performs so much better on the test set but so much worse on the cross validation set.

Confusion matrices were made for the top three test set scoring models: downsampling on logistic regression, no change on logistic regression, and downsampling on gradient boost. As seen in Figure 10, Figure 11, and Figure 12, all models suffer from poor precision in predicting a cell as cancerous. Downsampling on logistic regression had the highest cancerous cell class precision score of 0.38 while the other two had scores of 0.35. The downsampling on gradient boost model was able to correctly predict the highest number of cells as cancerous out of the three.

5.5. A ∪ B vs. A ∩ B on the Downsampled Dataset

To predict cancer cell types, the features of the model are from the marker genes of Cancer Clone A and the marker genes of Cancer Clone B. The effect of using the union of the set of marker genes from Cancer Clone A and B versus the effect of using the intersection of marker genes from Cancer Clone A and B was studied. A ∪ B resulted in 156 marker genes as features of the model whereas A ∩ B had 34 genes (Figure 3). The dataset was also balanced with the same method of downsampling from section 5.2 to reduce the number of noncancerous cells. The data was also split into the same 80% for cross validation training and same 20% for testing like previously. Gridsearch 5-fold cross validation with f1 scoring was used to tune the hyperparameters of all four models (logistic regression, gradient boost, random forest, support vector machine) as well.

With gridsearch 5-fold cross validation, the validation f1 scores and error bars from all four models were graphed (Figure 13). Although the error bars for 3 out of the 4 models do not overlap, there is a noticeable trend that taking the A ∩ B as gene features for the model scores better than A ∪ B.

Based on the length of error bars obtained from gridsearch cross validation in Figure 13, A ∪ B on logistic regression has the most variability. The other models with either union or intersection have similar f1 score spreads close to each other.

Using the best parameters found via gridsearch 5-fold cross validation, the entire 80% training dataset was fit onto the model and predicted on 20% test dataset. The bar graph of the test set f1 scores in Figure 13 show that taking the intersection of marker genes from Cancer Clone A and B performs better than taking the union of them. On the test set, gene feature selection by A ∩ B on logistic regression performs the best, with a f1 score of 0.85. The next top performing model is A ∩ B on support vector machine, scoring a f1 score of 0.74. The models and parameters that performed the best are the same ones from section 5.3.

5.6. A ∪ B vs. A ∩ B on the No Change Dataset

Once again, the same process from section 5.5 is run but on the nondownsampled dataset. F1 scores of the cross validation set and error bars are generated for each of the four models. The error bars do not overlap for logistic regression, random forest, and SVM, suggesting that using the intersection of marker genes from cancer cluster A and B on the nondownsampled dataset is statistically significantly better than the union of marker genes from the two clusters.

On the test set, a similar trend is observed. The intersection of marker genes performs better than the union for each model. The best performing model on the test set was using A ∩ B marker genes as features for a logistic regression, where a f1 score of 0.66 was achieved. It is interesting to note that both logistic regression models, whether it be A ∪ B or A ∩ B for features, have very low seen variability (based on the length of error bars in Figure 14) compared to the other models. Both logistic regression models also perform the best on the test set compared to any other model.

Comparing the confusion matrix of A ∩ B on logistic regression in Figure 15 to the confusion matrix of A ∪ B on logistic regression in Figure 16, it seems that the A ∩ B feature logistic regression model can classify true positive cancerous cells better.

Through Figure 17, it can be seen that for each of the 4 experiments- Downsampling Vs. No Change on A ∩ B, Downsampling Vs. No Change on A ∪ B, Downsampling Vs. No Change on A ∪ B, and Downsampling Vs. No Change on A ∩ B- the logistic regression model performs the best on the test set. Overall, between all 16 different settings ran, the top three performing settings were A ∩ B with downsampling on logistic regression (0.85), A ∩ B with downsampling on support vector machine (0.74), and A ∩ B with no downsampling on logistic regression (0.66).

Based on the results, using A ∩ B as the model's gene features performed better than A ∪ B. The intersection of these genes will give the set of marker genes that are important to the expression of both Cancer Clone A cell type and B cell type. On the other hand, the union contains a set of genes

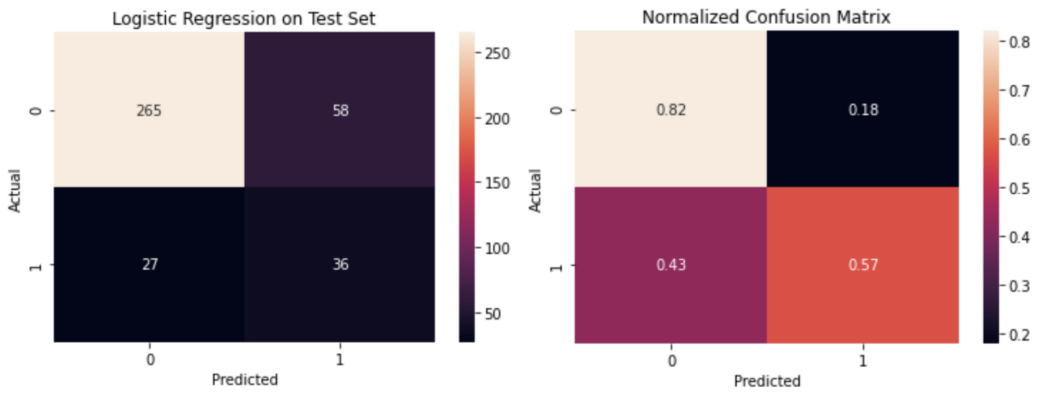

Figure 10. Confusion matrix and normalized confusion matrix of test set predictions using A ∪ B, downsampling on logistic regression.

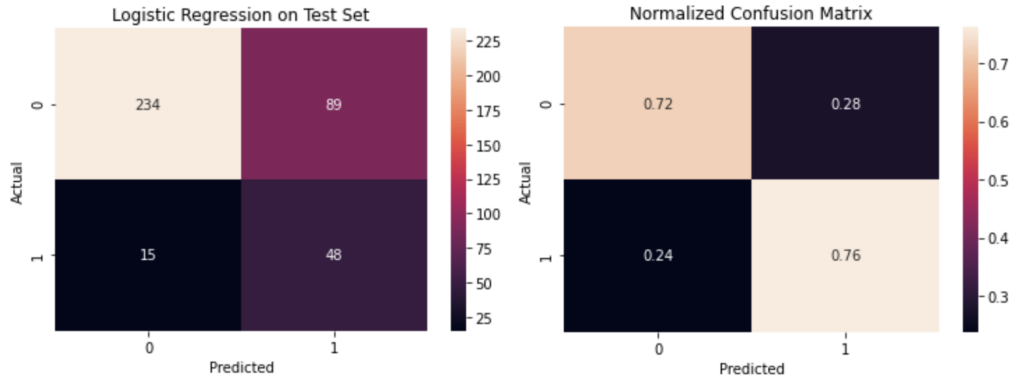

Figure 11. Confusion matrix and normalized confusion matrix of test set predictions using A ∪ B, no downsampling on logistic regression.

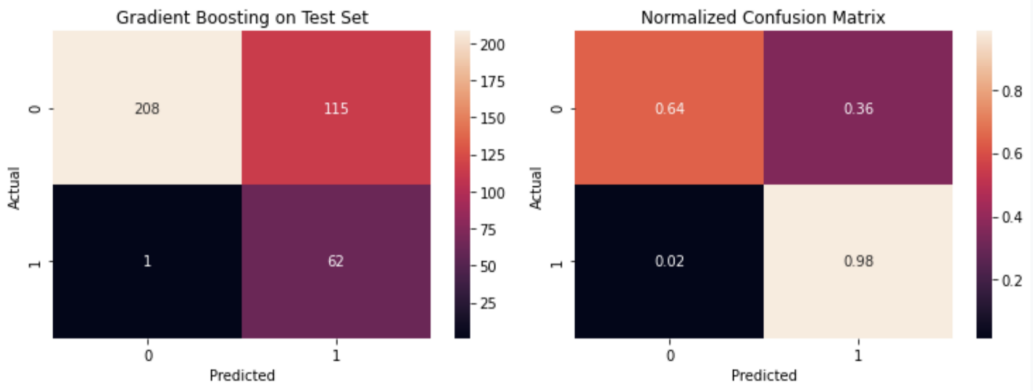

Figure 12. Confusion matrix and normalized confusion matrix of test set predictions using A ∪ B, downsampling on gradient boost.

important to the expression of A but potentially not B, and vice versa.

Downsampling the majoirty class label of the dataset also performed better than not changing anything. This behavior can be rationalized because downsampling a heavily class imbalanced dataset helps prevent the model from developing a bias towards the majority class.

5.7. Feature Importance

To determine what specific genes are helping make the models' predictions the most, feature importance scores were calculated for the downsampled A ∩ B on logistic regression and random forest models.

For logistic regression, the coefficient determined from the

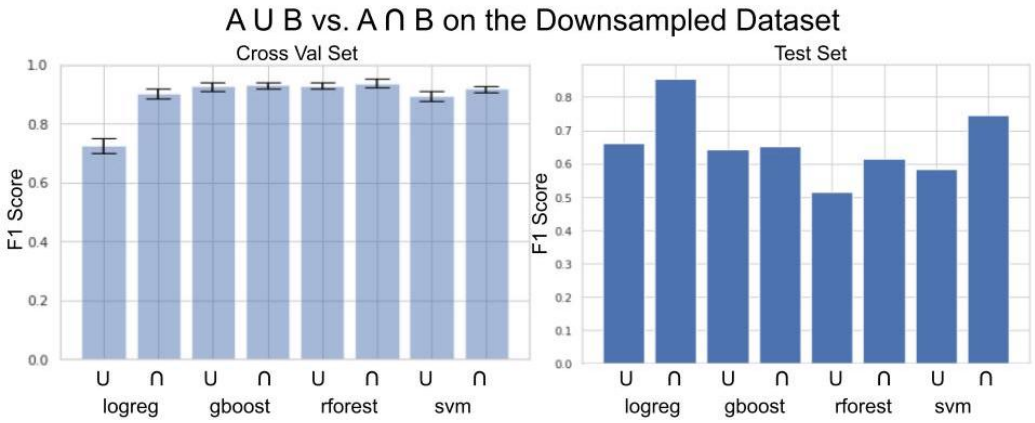

Figure 13. Bar graph of the f1 scores retrieved from four different machine learning models when using A ∪ B versus A ∩ B as gene features. The dataset used to train the model is downsampled to combat class imbalance too.

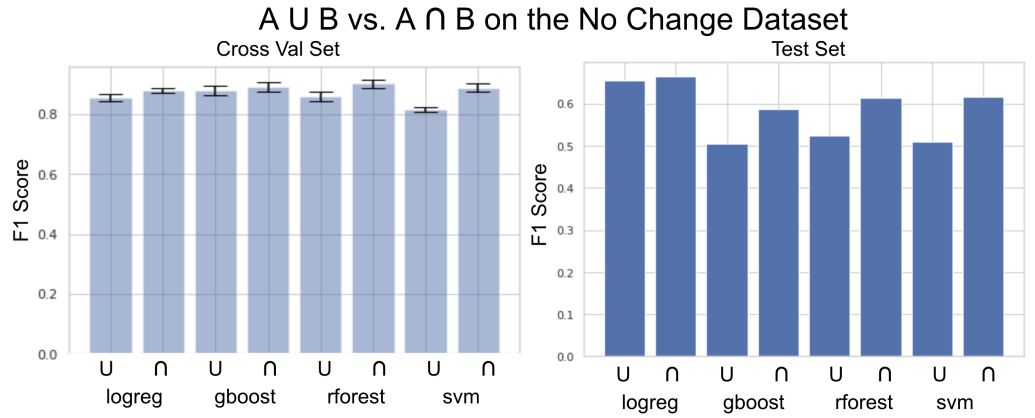

Figure 14. Bar graph of the f1 scores retrieved from four different machine learning models when using A ∪ B versus A ∩ B as gene features. The dataset used to train the model is not downsampled, and the class imbalance is left in.

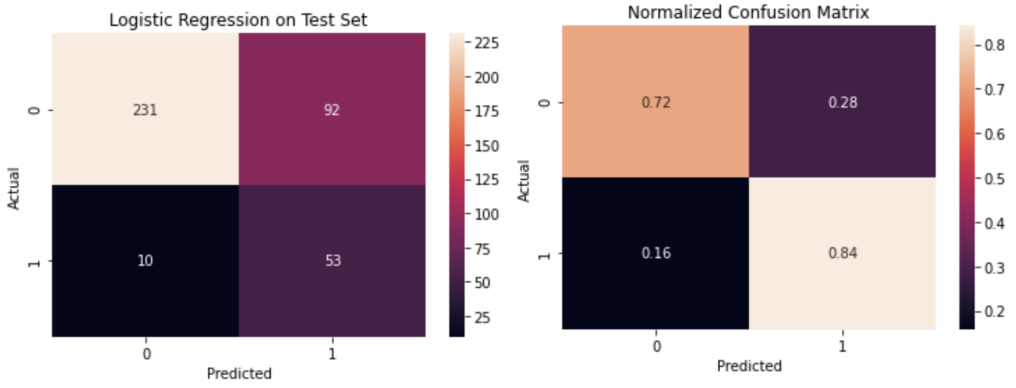

Figure 15. Confusion matrix and normalized confusion matrix of test set predictions using A ∩ B, no downsampling on logistic regression.

model can provide the basis for a crude feature importance score (Brownlee, 2020). Since single cell RNA sequencing data is very heavily zero inflated, it was first dimension reduced with PCA before inputting into any of the models. Hence, any feature importance extracted from the model is based on the PCA components.

In the logistic regression feature importance graph, it can be

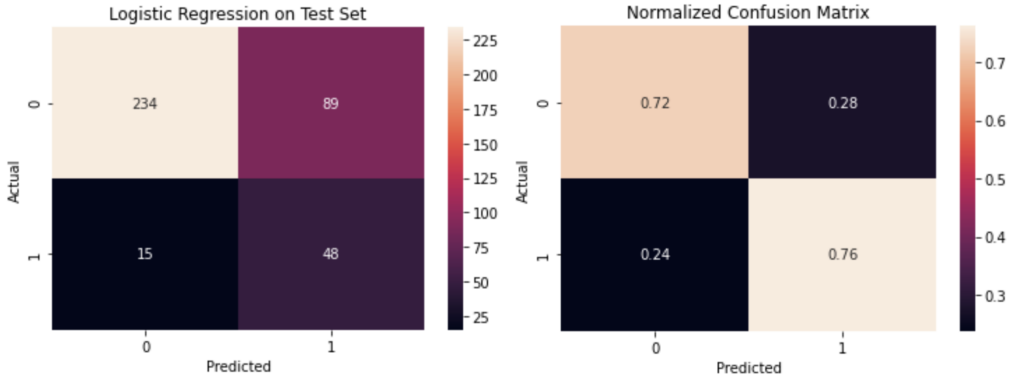

Figure 16. Confusion matrix and normalized confusion matrix of test set predictions using A ∪ B, no downsampling on logistic regression.

F1 Scores on Test Set

	Model	A ∪ B	A ∩ B
No Change	Logistic Regression	**0.65**	**0.66**†
	Gradient Boosting	0.50	0.59
	Random Forest	0.52	0.61
	Support Vector Machine	0.51	0.62
Down Sampling Majority	Logistic Regression	**0.66**	**0.85**†
	Gradient Boosting	0.64	0.65
	Random Forest	0.51	0.61
	Support Vector Machine	0.58	0.74†

Figure 17. Table of f1 scores of the various 16 settings and models' predictions on the test set. Scores in bold represents the best out of each of the four main settings. A † superscript represents the best three scores out of all 16.

seen that PC1, PC2, and PC3 have the largest absolute value coefficients. They have the most importance to the model when predicting a label. Since the first and second principle component are both positive, they are features that predict class 1. On the other hand, since the third component is negative, this feature contributes to predicting class 0.

The feature importance of genes from the random forest model were determined by running the random forest 1000 times. The feature importance of the genes is graphed with a violin plot. In the random forest feature importance graph, it can seen once again that the first three principal components, PC1, PC2, and PC3, have the highest feature importance. They have the most importance on any prediction that comes out of the random forest model. PC11 and PC12 also stand out in Figure 19 as important features for the random forest model.

The fact that the first three principal components for logistic regression and random forest are the ones most important to the model when making a prediction goes hand in hand with the background knowledge of PCA. This is since in PCA, the first component explains the most variation and each successive PCA component explains less and less.

Then, the genes that make up the most of each principal component was analyzed. From analyzing the feature importance from Figure 18 and Figure 19, a table of principle components and their corresponding top contributing gene was constructed. From the collection of genes in Figure 20, the scVI + UMAP graph was colored by gene expression to see which cell types these genes were highly expressed in.

Looking at the colored gene expression, it can be seen that the determined genes do have high expression in the Cancer Clone A and B clusters. Genes KRT17 and S100P have some high expression in the ductal cell clusters too. It is interesting to note that the two genes from PC11 and PC12, PTGS2 and COL17A1, are highly expressed in cancerous cluster A and B and very lowly expressed everywhere

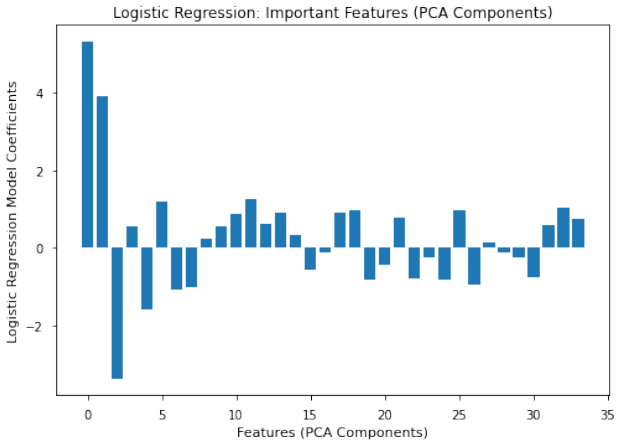

Figure 18. Bar chart of logistic regression coefficients as feature importance scores. The logistic regression model used A ∩ B as gene features and a downsampled majority class dataset.

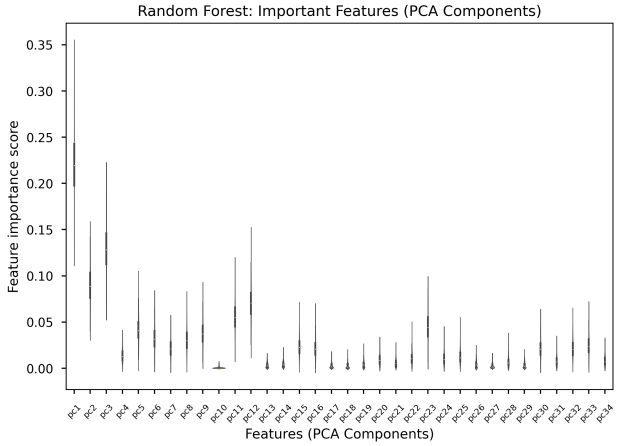

Figure 19. Violin plot of random forest feature importance scores. The random forest model used A ∩ B as gene features and a downsampled majority class dataset.

PC	Gene
PC1	KRT17
PC2	S100A4
PC3	S100P
PC11	PTGS2
PC12	COL17A1

Figure 20. Table of the top contributing gene for five notable principle components of PCA: PC1, PC2, PC3, PC11, PC12.

else. Since the exact composition of tumor subclassification varies from person to person, the cells subpopulation composition and marker gene identification can be analyzed for each given patient and be of prognostic use.

5.7.1. Results Validation

The "association" between the model's top contributing genes and pancreatic cancer were validated through literature found by PubMed (https://www.ncbi.nlm.nih.gov/pmc/).

Publications have found that the overexpression of KRT17 results in shorter patient survival (Pan et al., 2020; Zeng et al., 2020). Literature has also shown that the gene regulates the proliferation, migration, and invasion of pancreatic cancer cells (Li et al., 2020; Chen et al., 2020). Another study by Pan et al. reported that KRT17 has the potential to be a novel target for making biomarker-based personalized treatment for PDAC (Pan et al., 2020).

As for the gene S100A4, it is known to be associated with poor clinical outcomes for patients with pancreatic cancer. Jia et al. discovered that when combined with serum CA19.9, it serves as dual prognostic biomarkers for pancreatic cancer (Jia et al., 2019). Another study by Che et al. found that S100A4 is important to the progression of pancreatic cancer in vivo and promotes tumorigenic phenotypes of pancreatic cancer cells (Che et al., 2015).

In literature, analysis of S100P showed high promise in the treatment and diagnosis of pancreatic adenocarcinoma (Chiba et al., 2017; Dakhel et al., 2014). Camara et al. also establishes the potential in using S100P's small molecule inhibitors for drug candidate development (Camara et al., 2020).

For the gene PTGS2, it is known to be a key factor in inflammation, and its expression is heightened in pancreatic cancer (Wei et al., 2004). High levels of PTGS2 is also associated with poor prognosis of cancer (Ristimäki et al., 2002; Kondo et al., 1999; Masunaga et al., 2000). When combined with the transcription factor SP1 to form a signaling axis, PTGS2 has shown to play an important role in pancreatic cancer (Bennett et al., 2016).

In literature, high COL17A1 expression was also seen to be associated with a poorer overall survival. The gene was found to have important prognostic value and potential as a molecular therapeutic target for PDAC treatment (Mao et al., 2020). When analyzing survival in patients with pancreatic cancer, COL17A1 was also established as one of the important gene expression signatures (Zhang et al., 2020; Wu et al., 2019).

6. Conclusion

In this study, we present a cell level and gene level analysis of pancreatic ductal adenocarcinoma (PDAC) tissue. By using the latent space to cluster cell types, we demonstrated

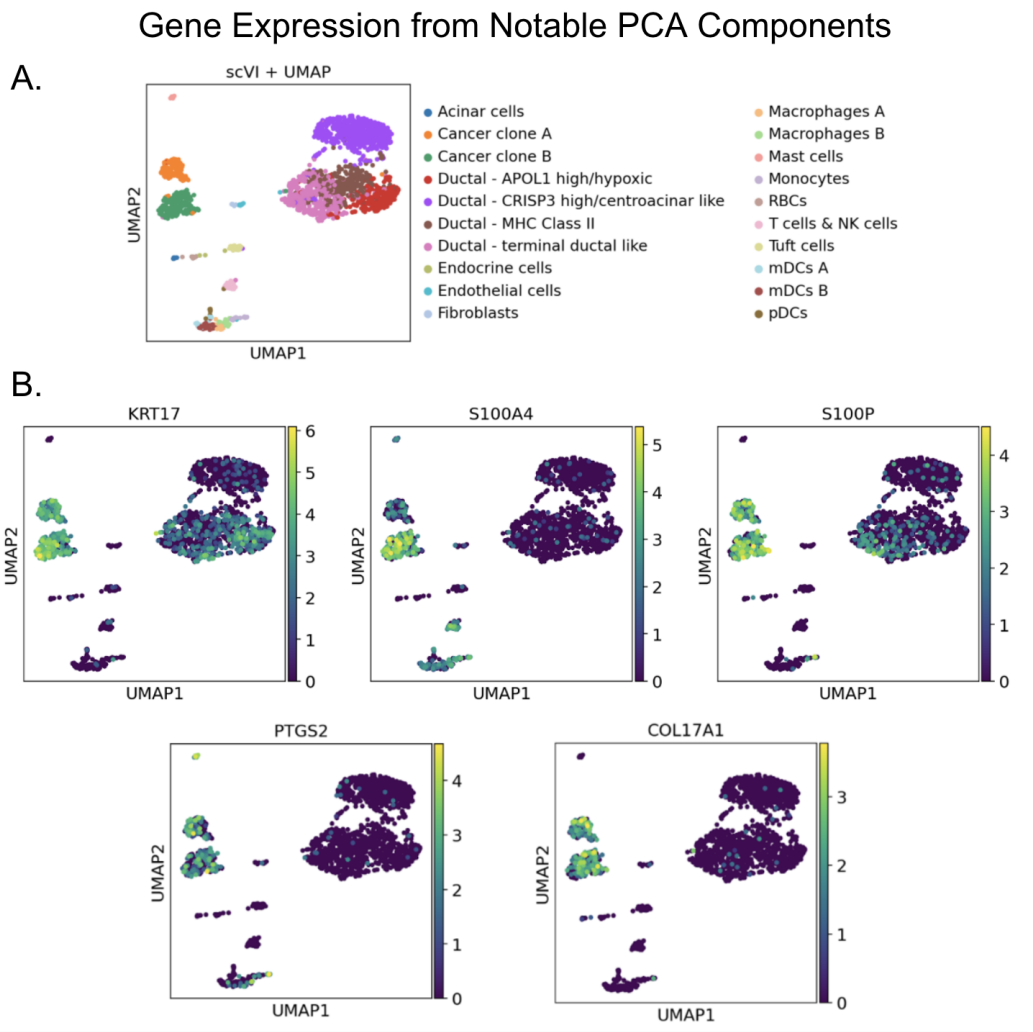

Figure 21. A. Clustering of cell types with UMAP on top of scVI latent dimension mapping B. The same cell clustering map from part A but now colored by gene expression. The genes used to color were the top contributing gene for five notable principle components of PCA: PC1, PC2, PC3, PC11, PC12

that this method worked better at resolving cell resolution and is a high performing alternative to PCA dimension reduction. Then, with marker genes determined from cancerous cell types by differential expression, classification models were used to predict cancerous cells. Different settings were experimented to combat data imbalance of the cancerous class and to determine the best set of marker gene features for the models. Downsampling the majority class and using the intersection of marker genes from cancer clone A and B were determined to give the models the best f1 scoring. Because of downsampling, the best performing model, logistic regression, was able to improve from a f1 score of 0.664 to 0.854. By fitting the model to only the intersection of cancer cluster A and B's marker genes instead of the union, the logistic regression model's f1 score improved from 0.660 to 0.854. By tracing back the top contributing genes of the classification models' feature importance, the top genes were then analyzed by cell type expression and validated through literature. This analysis of gene expression and cell subpopulations of PDAC tissue can allow for a more precise understanding of the transcriptome in cells at an individual level. The functions and underlying cellular mechanisms of pancreatic ductal adenocarcinoma are also able to be uncovered. When analyzed for a given patient's tumor tissue, it can be of prognostic value as well.

7. Future Work

7.1. Clustering Metrics

The results of clustering with PCA or scVI were visually analyzed to determine cell resolution and accuracy of cell sub-

population identification. A metric could be implemented to quantify the accuracy of the results and would make comparing the two clustering methods more robust.

7.2. Data Imbalance

The best tested method of combating the heavy data imbalance was downsampling the majority class, the number of noncancerous cells. Another method that can be experimented with is upsampling by generating new data, such as with SMOTE (Chawla et al., 2002) or GANs (Goodfellow et al., 2014).

7.3. Spatial Transcriptomics

Supplementary information from my dataset source also included spatial transcriptomics, an emerging field of study that analyzes the spatial architecture and cell location of tissue. A future area of work would be combining and analyzing this spatial transcriptomics data with the single cell RNA sequencing to uncover the structural organization of the tumour.

References

SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis.

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., and Yang, G.-Z. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.

Auer, P. L. and Doerge, R. Statistical design and analysis of rna sequencing data. *Genetics*, 185(2):405–416, 2010.

Bennett, C. L., Lai, S. Y., Sartor, O., Georgantopoulos, P., Hrushesky, W. J., Henke, M., and Armitage, J. O. Consensus on the existence of functional erythropoietin receptors on cancer cells. *JAMA oncology*, 2(1):134–136, 2016.

Briggs, J. A., Weinreb, C., Wagner, D. E., Megason, S., Peshkin, L., Kirschner, M. W., and Klein, A. M. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science (New York, N.Y.)*, 360 (6392):eaar5780, June 2018. ISSN 0036-8075. doi: 10.1126/science.aar5780.

Brownlee, J. How to Calculate Feature Importance With Python, March 2020.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33 (2):155–160, 2015.

Camara, R., Ogbeni, D., Gerstmann, L., Ostovar, M., Hurer, E., Scott, M., Mahmoud, N. G., Radon, T., Crnogorac-Jurcevic, T., Patel, P., et al. Discovery of novel small molecule inhibitors of s100p with in vitro anti-metastatic effects on pancreatic cancer cells. *European journal of medicinal chemistry*, 203:112621, 2020.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. arXiv: 1106.1813.

Che, P., Yang, Y., Han, X., Hu, M., Sellers, J. C., Londono-Joshi, A. I., Cai, G.-Q., Buchsbaum, D. J., Christein, J. D., Tang, Q., et al. S100a4 promotes pancreatic cancer progression through a dual signaling pathway mediated by src and focal adhesion kinase. *Scientific reports*, 5(1): 1–10, 2015.

Chen, P., Shen, Z., Fang, X., Wang, G., Wang, X., Wang, J., and Xi, S. Silencing of keratin 17 by lentivirus-mediated short hairpin rna inhibits the proliferation of panc-1 human pancreatic cancer cells. *Oncology letters*, 19(5): 3531–3541, 2020.

Chiba, M., Imazu, H., Kato, M., Ikeda, K., Arakawa, H., Kato, T., Sumiyama, K., and Homma, S. Novel quantitative analysis of the s100p protein combined with endoscopic ultrasound-guided fine needle aspiration cytology in the diagnosis of pancreatic adenocarcinoma. *Oncology reports*, 37(4):1943–1952, 2017.

Dakhel, S., Padilla, L., Adan, J., Masa, M., Martinez, J., Roque, L., Coll, T., Hervas, R., Calvis, C., Messeguer, R., et al. S100p antibody-mediated therapy as a new promising strategy for the treatment of pancreatic cancer. *Oncogenesis*, 3(3):e92–e92, 2014.

Delmans, M. and Hemberg, M. Discrete distributional differential expression (D3E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics*, 17(1):110, February 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0944-6.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., and Gottardo, R. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, December 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0844-5.

Gaublomme, J. T., Yosef, N., Lee, Y., Gertner, R. S., Yang, L. V., Wu, C., Pandolfi, P. P., Mak, T., Satija, R., Shalek,

A. K., et al. Single-cell genomics unveils critical regulators of th17 cell pathogenicity. *Cell*, 163(6):1400–1412, 2015.

Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N., and Garry, D. J. Drimpute: imputing dropout events in single cell rna sequencing data. *BMC bioinformatics*, 19(1): 1–10, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

Guo, M., Wang, H., Potter, S. S., Whitsett, J. A., and Xu, Y. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, 9(1):1–12, 2017.

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1):29, February 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0888-1.

Jia, F., Liu, M., Li, X., Zhang, F., Yue, S., and Liu, J. Relationship between s100a4 protein expression and preoperative serum ca19. 9 levels in pancreatic carcinoma and its prognostic significance. *World journal of surgical oncology*, 17(1):1–10, 2019.

Khalfaoui, B. and Vert, J.-P. DropLasso: A robust variant of Lasso for single cell RNA-seq data. *arXiv:1802.09381 [cs, q-bio, stat]*, February 2018.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014. ISSN 1548-7105. doi: 10.1038/nmeth.2967. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genome-wide analysis of gene expression;Statistical methods;Transcriptomics Subject_term_id: genome-wide-analysis-of-gene-expression;statistical-methods;transcriptomics.

Kondo, M., Yamamoto, H., Nagano, H., Okami, J., Ito, Y., Shimizu, J., Eguchi, H., Miyamoto, A., Dono, K., Umeshita, K., et al. Increased expression of cox-2 in nontumor liver tissue is associated with shorter disease-free survival in patients with hepatocellular carcinoma. *Clinical cancer research*, 5(12):4005–4012, 1999.

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology*, 17(1):222, October 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-1077-y.

Li, D., Ni, X.-F., Tang, H., Zhang, J., Zheng, C., Lin, J., Wang, C., Sun, L., and Chen, B. Krt17 functions as a tumor promoter and regulates proliferation, migration and invasion in pancreatic cancer via mtor/s6k1 pathway. *Cancer management and research*, 12:2087, 2020.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0229-2.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Mao, F., Li, D., Xin, Z., Du, Y., Wang, X., Xu, P., Li, Z., Qian, J., and Yao, J. High expression of col17a1 predicts poor prognosis and promotes the tumor progression via nf-κb pathway in pancreatic adenocarcinoma. *Journal of Oncology*, 2020, 2020.

Masunaga, R., Kohno, H., Dhar, D. K., Ohno, S., Shibakita, M., Kinugasa, S., Yoshimura, H., Tachibana, M., Kubota, H., and Nagasue, N. Cyclooxygenase-2 expression correlates with tumor neovascularization and prognosis in human colorectal carcinoma patients. *Clinical Cancer Research*, 6(10):4064–4068, 2000.

Miao, Z., Deng, K., Wang, X., and Zhang, X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34(18): 3223–3224, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty332.

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., Hajdu, C. H., Simeone, D. M., and Yanai, I. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology*, 38(3):333–342, March 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0392-8.

Müller, S., Cho, A., Liu, S. J., Lim, D. A., and Diaz, A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics*, 34(18):3217–3219, September 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty316.

Pan, C.-H., Otsuka, Y., Sridharan, B., Woo, M., Leiton, C. V., Babu, S., Torrente Gonçalves, M., Kawalerski, R. R., K. Bai, J. D., Chang, D. K., et al. An unbiased high-throughput drug screen reveals a potential therapeutic vulnerability in the most lethal molecular subtype of pancreatic cancer. *Molecular oncology*, 14(8):1800–1816, 2020.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

Poirion, O. B., Zhu, X., Ching, T., and Garmire, L. Single-cell transcriptomics bioinformatics and computational challenges. *Frontiers in genetics*, 7:163, 2016.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. Single-cell mRNA quantification and differential analysis with Census. *Nature Methods*, 14(3):309–315, March 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4150. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biotechnology;Computational biology and bioinformatics Subject_term_id: biotechnology;computational-biology-and-bioinformatics.

Ristimäki, A., Sivula, A., Lundin, J., Lundin, M., Salminen, T., Haglund, C., Joensuu, H., and Isola, J. Prognostic significance of elevated cyclooxygenase-2 expression in breast cancer. *Cancer research*, 62(3):632–635, 2002.

Scialdone, A., Natarajan, K. N., Saraiva, L. R., Proserpio, V., Teichmann, S. A., Stegle, O., Marioni, J. C., and Buettner, F. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

Semrau, S., Goldmann, J. E., Soumillon, M., Mikkelsen, T. S., Jaenisch, R., and Van Oudenaarden, A. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature communications*, 8(1):1–16, 2017.

Stegle, O., Teichmann, S. A., and Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

Vrahatis, A. G., Tasoulis, S. K., Maglogiannis, I., and Plagianakos, V. P. Recent Machine Learning Approaches for Single-Cell RNA-seq Data Analysis. In Maglogiannis, I., Brahnam, S., and Jain, L. C. (eds.), *Advanced Computational Intelligence in Healthcare-7: Biomedical Informatics*, Studies in Computational Intelligence, pp. 65–79. Springer, Berlin, Heidelberg, 2020. ISBN 978-3-662-61114-2. doi: 10.1007/978-3-662-61114-2_5.

Wang, T. and Nabavi, S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods*, 145:25–32, August 2018. ISSN 1046-2023. doi: 10.1016/j.ymeth.2018.04.017.

Wei, D., Wang, L., He, Y., Xiong, H. Q., Abbruzzese, J. L., and Xie, K. Celecoxib inhibits vascular endothelial growth factor expression in and reduces angiogenesis and metastasis of human pancreatic cancer via suppression of sp1 transcription factor activity. *Cancer research*, 64(6):2030–2038, 2004.

Wolf, F. A., Angerer, P., and Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

Wu, M., Li, X., Zhang, T., Liu, Z., and Zhao, Y. Identification of a nine-gene signature and establishment of a prognostic nomogram predicting overall survival of pancreatic cancer. *Frontiers in oncology*, 9:996, 2019.

Zeng, Y., Zou, M., Liu, Y., Que, K., Wang, Y., Liu, C., Gong, J., and You, Y. Keratin 17 suppresses cell proliferation and epithelial-mesenchymal transition in pancreatic cancer. *Frontiers in Medicine*, 7:847, 2020.

Zhang, Z., Gu, J., Yin, M., Wang, D., Ma, C., Du, J., Lin, Z., Hu, S., Wang, X., Li, Y., et al. Establishment and investigation of a multiple gene expression signature to predict long-term survival in pancreatic cancer. *BioMed research international*, 2020, 2020.