Identification of Cancer Cells at a Single-cell Level

December 19, 2022

1 Introduction

Tumors are complex tissues made up of malignant cells that are surrounded by a varied cellular milieu with which they interact. The molecular characterization of individual tumor cells is made possible by single-cell sequencing. The task of cell annotation, which entails assigning a cell type or condition to each sequenced cell, is difficult, particularly when trying to detect tumor cells in single-cell or spatial sequencing research. The purpose of this project is to use multiple machine learning approaches to perform single-cell analysis to identify cancer cells in single-cell RNA sequencing (scRNA) data. Specifically, we are hoping to distinguish tumor cells from normal cells of colorectal cancer (CRC) at the single-cell level. As increasingly more scRNA data become available, manually selecting a gene marker set for classification model will be time-consuming and repetitive. Dohmen et al., identified tumor cells at single-cell level using machine learning by developing a machine learning pipeline (Dohmen et al., 2022) and a review paper by Asada et al., showed the comprehensive applications using machine learning in single-cell analysis and its application to medical research (Asada et al., 2021). Because cancers are caused by changes to the systems of life, understanding life at the single-cell level is crucial for medical research to understand and control them. Particularly, cancer cells constitute different mutations and characteristics, analysis at single-cell level is especially essential to understand cancer tissue diversity to contribute to the establishment of potential cancer therapies. In brief, we built three supervised machine learning models that are able to differentiate cancer and normal cells in colorectal cancer single-cell RNA seq data. We also explored how to process Single-Cell RNA-seq data step by step.

2 Data

We used scRNA sequencing data of colorectal cancer from NCBI Gene Expression Omnibus public database. We chose these datasets because they are all raw scRNA-seq data from CRC patients with colorectal cancer samples as well as normal samples without any additional treatments. Thus, we can use them for training and testing our models without any bias. These datasets include:

1. GSE200997: droplet-based scRNA-seq on 16 racially diverse, treatment naïve CRC patient tissue samples and seven adjacent normal colonic tissue samples, yielded $49,589$ single cells including tumor and microenvironmental cells.

2. GSE144735: single-cell 3' RNA sequencing data on $27,414$ cells from 6 CRC patients in core and border tumor regions, as well as in matched normal mucosa samples.

3. GSE132465: single-cell 3' RNA sequencing data on $63,689$ cells from 23 CRC patients with 23 primary colorectal cancer and 10 matched normal mucosa samples.

Datasets	Normal Samples	Tumor Samples	Features
1 (GSE144735)	9,736	8,254	33,694
2 (GSE200997)	18,273	31,586	23,828
3 (GSE132465)	16,404	47,285	33,694

Table 1: Datasets Description

3 Methods

3.1 Data Preprocessing

For each scRNA-seq dataset, we performed quality control process to filter the low quality cells that have unique feature counts over 6,000 or less than 200, and also filter the possible dying cells that have over 20% mitochondrial counts. We also discarded the features that are rarely detected and only keep the features detected in at least 3 cells. For the normalization step, feature counts for each cell are divided by the total counts for that cell and multiplied by the scale factor and then natural-log transformed. We next calculate a subset of features that exhibit high cell-to-cell variation in the dataset (i.e, they are highly expressed in some cells, and lowly expressed in others), since it is found that focusing on these genes in downstream analysis helps to highlight biological signal in single-cell datasets. By modeling the mean-variance relationship inherent in single-cell data, the 3,000 most variable genes were selected per dataset as the features for downstream analysis like PCA and clustering. We also implemented PCA to conduct dimensionality reduction and visualize the scRNA-seq datasets. The scRNA-seq data preprocessing was implemented in R by utilizing the public package Seurat.

3.2 Clustering and Identification of Marker Genes

To better visualize the clustering of tumor cells and normal cells, we implemented the non-linear dimensional reduction UMAP to explore the clusters in low-dimensional space, using the top 10 PCs generated in PCA. Cells from the same cell type should co-localize on these dimension reduction plots. To identify a set of marker genes which can be used as features for classifying tumor cells from normal cells, we identified the significantly differentially expressed genes between cancer cells and normal cells for all datasets with a log fold change cutoff = ± 0.5 and an adjusted p-value threshold = 0.05. In cases of multiple testing, p-values are adjusted using Benjamini-Hochberg (FDR) method. The differential expression analysis was implemented by the public edgeR package.

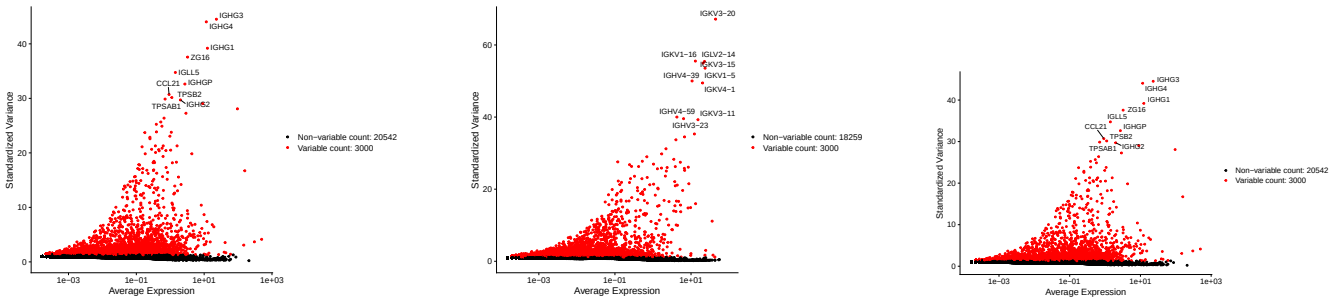
3.3 Classification and Evaluation

Then, we used identified marker genes as input features and applied the supervised learning by training several robust classifiers including logistic regression, random forest, and support vector machine for stringent discrimination of tumor and normal cells using sklearn package from Python for each dataset.
In order to analyze the results, we computed the cross-validation using k-fold cross-validation methods for each method and each dataset and generated the receiver operating characteristic curve which is a graph showing the performance of a classification model at all classification thresholds for the results.

4 Results

4.1 Data Preprocessing

For GSE144735 dataset, we filtered 1 low quality cells and 10152 rare features; For GSE200997 dataset, we filtered 606 low quality cells and 2569 rare features; For GSE132465 dataset, we filtered 102 low quality cells and 9995 rare features. The variance-mean relationship for 3 datasetes is shown as figure 1, and we selected 3,000 most variable genes for following analysis.

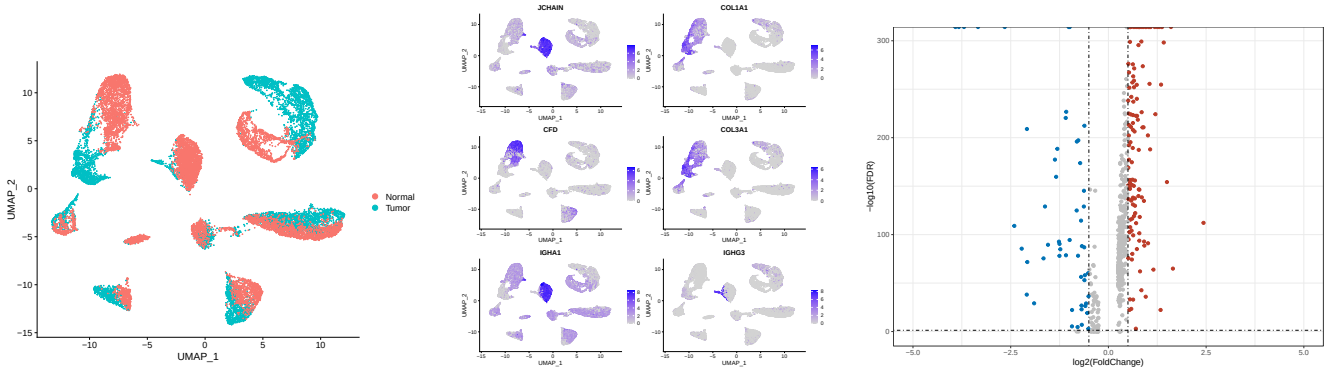


(a) dataset 1 (b) dataset 2 (c) dataset 3

Figure 1: Selected Features for Three Datasets.

4.2 Clustering and Identification of Marker Genes

By applying non-linear dimensional reduction, the tumor cells and normal cells are separated into different clusters as shown in figure 2(a). The marker genes for tumor cells and normal cells are identified as the differentially expression genes between the two cell types. The tumor marker genes are highly expressed in the tumor clusters and the normal marker genes are also enriched in the clusters which are overlapped with the normal clusters, shown as figure 2(b). We identified 549 DEGs in total, including significantly down-regulated as well as up-regulated genes, shown as the volcano plot in figure 2(c).



(a) UMAP (b) Identified Marker Genes (c) DEGs

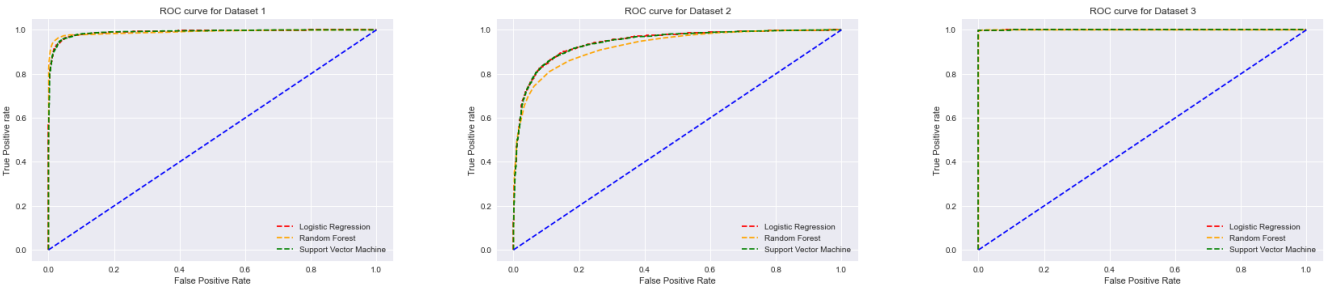
Figure 2: Clustering and Identification of Marker Genes

4.3 Classification

The datasets were randomly sampled to training sets and testing sets for training and testing each model and the accuracy scores were calculated for each method (Table 2). For dataset 1, random forest exhibits the highest accuracy score of 0.967 while support vector machine gives the highest accuracy score in both dataset 2 and dataset 3. Overall, the three classification models gave a high accuracy scores in classifying all three datasets with accuracy scores around 0.9. We also plot the ROC curves for each model using the `roc_curve` function from sklearn.metrics which would automatically generate thresholds based on predicted probabilities and $n\_thresholds = len(np.unique(y\_pred)) + 1$ (Figure 1). As expected, the ROC curves all exhibited great performances for all classification methods.

Methods	Logistic Regression	Random Forest	Support Vector Machine
1 (GSE144735)	0.955	0.973	0.960
2 (GSE200997)	0.874	0.852	0.875
3 (GSE132465)	0.960	0.870	1.000

Table 2: Accuracy Scores from Classification Models for All Three Datasets

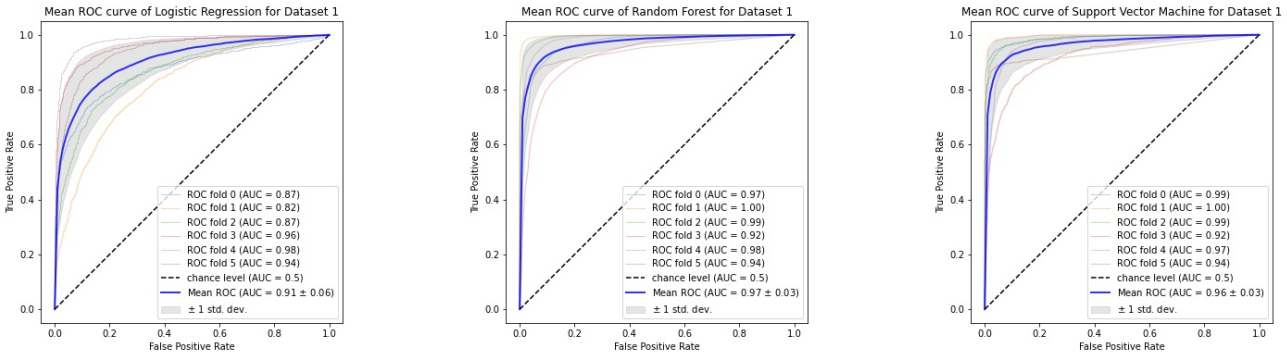


(a) dataset 1 (b) dataset 2 (c) dataset 3

Figure 3: ROC Curves of Logistic Regression, Random Forest and Support Vector Machine for Three Datasets.

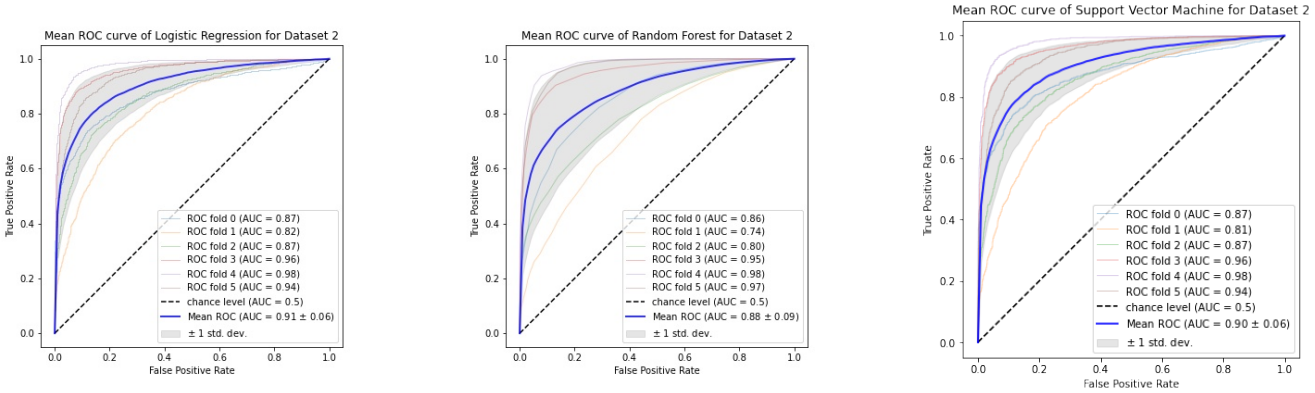
4.4 Evaluation

A typical way to evaluate a machine learning model is to divide data set into training and testing sets, with the training set being used to train the model and the testing set being used to test the model. The model's correctness is then determined by evaluating its performance using an error measure. This approach, on the other hand, is not very dependable since the accuracy gained for one test set might be substantially different from that obtained for another test set. We can solve the issue by separating the data into stratified k-folds and each time we using different combinations of several of the parts as training set and testing set. We used the six-fold cross validation to validate the all three model results. Every time we used 80% of the data set as training set, and 20% of the data set as testing set. In the first iteration, we used the first fold data to test the model, and the rest of the data to train the model and then continued the training until each of the six folds has been utilized as a test set. The results for cross-validation are presented using ROC curves which are shown below. They all exhibited high AUC scores which validated our model can distinguish tumor cells from normal cells in scRNA-seq data.



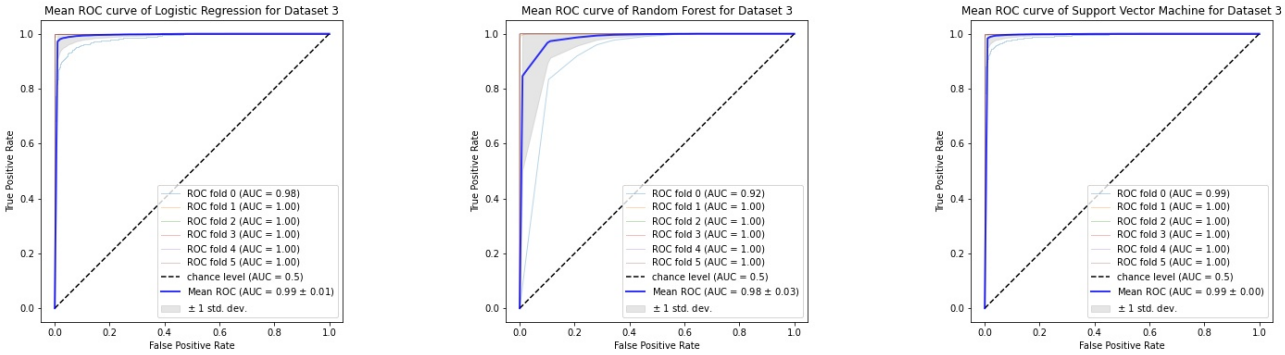
(a) Logistic Regression (b) Random Forest (c) Support Vector Machine

Figure 4: ROC curve and AUC of each fold of Machine Learning Classification Models for Dataset 1.



(a) Logistic Regression (b) Random Forest (c) Support Vector Machine

Figure 5: ROC curve and AUC of each fold of Machine Learning Classification Models for Dataset 2.



(a) Logistic Regression (b) Random Forest (c) Support Vector Machine

Figure 6: ROC curve and AUC of each fold of Machine Learning Classification Models for Dataset 3.

5 Conclusion

In this project, we explore the pipeline for differentiating the tumor and normal cells. We first perform data pre-processing that discard low-quality cells using feature counts and mitochondrial counts. After normalization, we select the top 3000 genes for further analysis. Next, we conduct differential gene expression analysis and select 549 marker genes. By applying different machine learning models using marker genes as features, we achieve high accuracy. By further using cross-validation, we verify that the model performance is largely consistent given different test-train splits.

5.1 Potential Additional Analysis: Variant Calling

In addition, as Dohemen et al., attempts to use Copy Number Variations (CNV) extracted from RNA data to boost model performance, and observed an drop on false positive rate. We also explored the viability of variant calling on scRNA-seq datasets to see if detected variants can be used for differentiating tumor and normal cell. Our workflow is shown as below in Figure 7:

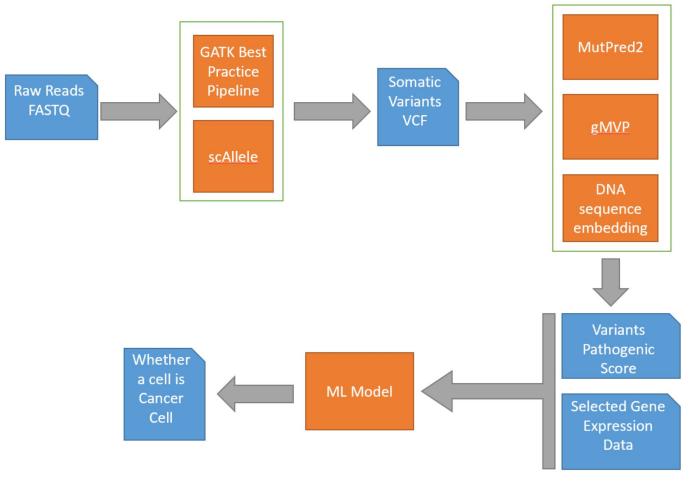


Figure 7: Workflow for using variant calling for differentiating tumor and normal cells. scAllele: tools that detect single-nucleotide variants in scRNA-seq; MutPred2: classifor amino acid substitutions as pathogenic or benign ; gMVP: a graphical missense variant pathogenicity predictor

However, we are not able to obtain the raw reads for the dataset we use for the previous analysis. And we use other raw read dataset that are not directly comparable, so we are not showing our results here. What we realize is that variant calling is extremely time-consuming, and it will be a topic that is worth studying but require a high volume of computing resources.

6 Reference

1. Asada, K., Takasawa, K., Machino, H., Takahashi, S., Shinkai, N., Bolatkan, A., Kobayashi, K., Komatsu, M., Kaneko, S., Okamoto, K., & Hamamoto, R. (2021). Single-Cell Analysis Using Machine Learning Techniques and Its Application to Medical Research. *Biomedicines*, 9(11), 1513.

2. Dohmen, J., Baranovskii, A., Ronen, J., Uyar, B., Franke, V., & Akalin, A. (2022). Identifying tumor cells at the single-cell level using machine learning. *Genome Biology*, 23(1).

3. Ronen, J., & Akalin, A. (2018). netSmooth: Network-smoothing based imputation for single cell RNA-seq. *F1000Research*, 7, 8.

4. Lee HO, Hong Y, Etlioglu HE, Cho YB et al. (2020). Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat Genet*, 52(6), 594-603.

5. Khaliq AM, Erdogan C, Kurt Z, Turgut SS et al. (2022). Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome Biol*, 23(1), 113.