Active feature selection discovers minimal gene-sets for classifying cell-types and disease states in single-cell mRNA-seq data

Xiaoqiao Chen[1], Sisi Chen[2,3], and Matt Thomson[1,2,3,4]

[1]Department of Computing and Mathematical Sciences, California Institute of Technology. Pasadena, California, 91125, USA.

[2]Division of Biology and Biological Engineering, California Institute of Technology. Pasadena, California, 91125, USA.

[3]Beckman Center for Single-cell Profiling and Engineering. Pasadena, California, 91125, USA.

[4]correspondence to: mthomson@caltech.edu

Abstract

Sequencing costs currently prohibit the application of single cell mRNA-seq for many biological and clinical tasks of interest. Here, we introduce an active learning framework that constructs compressed gene sets that enable high accuracy classification of cell-types and physiological states while analyzing a minimal number of gene transcripts. Our active feature selection procedure constructs gene sets through an iterative cell-type classification task where misclassified cells are examined at each round to identify maximally informative genes through an 'active' support vector machine (SVM) classifier. Our active SVM procedure automatically identifies gene sets that enables $> 90\%$ cell-type classification accuracy in the Tabula Muris mouse tissue survey as well as a ~ 40 gene set that enables classification of multiple myeloma patient samples with $> 95\%$ accuracy. Broadly, the discovery of compact but highly informative gene sets might enable drastic reductions in sequencing requirements for applications of single-cell mRNA-seq.

Introduction

Single-cell mRNA-seq methods have scaled to allow routine transcriptome-wide profiling of thousands of cells per experimental run. While single cell mRNA-seq approaches provide important insights into many different biological and biomedical problems, high sequencing costs prohibit the broad application of single-cell mRNA-seq in many exploratory assays such as drug and genetic screens as well as in cost sensitive clinical assays. The sequencing bottleneck has led to the development of targeted mRNA-seq strategies that reduce sequencing costs by focusing sequencing

resources on highly informative genes for a given biological question or an analysis [1, 2, 3, 4, 5]. Commercial gene targeting kits, for example, reduce sequencing costs through selective amplification of specific transcripts using 1000 gene targeting primers. Such approaches can reduce costs by 90% by focusing sequencing resources on highly informative transcripts.

Targeted sequencing approaches require computational methods to identify highly informative genes for specific biological questions, systems, or conditions. A range of computational approaches including differential gene expression analysis can be applied to identify highly informative genes. However, current methods for defining target gene sets often require heuristic user-defined thresholds for gene set selection and are computationally expensive to apply to large single cell mRNA-seq data sets [6, 7]. As an example computational approaches based upon matrix factorization (PCA[8], NNMF[9]) are typically applied to complete data sets which becomes prohibitively expensive when data sets scale into the millions of cells.

Here, we introduce a computational method that applies an active learning[10] approach to identify minimal gene sets for identifying cell-types and transcriptional states in single-cell mRNA-seq. Our method identifies minimal gene sets for performing a cell-type classification task of interest through iterative performance of classification followed by active examination of misclassified cells to discover highly-informative genes. Our method is inspired by Sequential Feature Selection (SFS)[11], where features (genes) are selected one by one. However, in our method, we approach sequential feature selection through an active learning strategy with a support vector machine[12] classifier ('active SVM'). At each round of iteration, our procedure analyzes only the cells that classify poorly with the current gene set, and the procedure extends the gene set by identifying genes within incorrectly classified cells that will maximally shift the classification margin. As the procedure focuses computational resources on poorly classified cells, the method can be applied to large data sets and discovers compact sets of genes that can distinguish between cell-types at high accuracy ($\sim 90\%$) from single-cell transcriptional data. Our active SVM methods, thus, provides an efficient computational strategy for constructing minimal gene sets that meet user defined classification goals from large single-cell mRNA-seq data sets.

We apply our active feature selection method to three test cases: a dataset of peripheral blood mononuclear cells (PBMCs) [13], the Tabula Muris mouse tissue survey[14], and a data set of multiple myeloma patient PBMCs [15]. We also systematically compare the performance of the method to six existing conventional feature selection methods, showing that our method outperforms other methods in terms of classification accuracy. Gene sets constructed by active SVM are extremely compact, and, for example, can by applied to efficiently classify human immune cell types within PMBCs using as few as 15 genes. The gene sets we discover include both classical markers as well as new genes not previously established as canonical markers. Our approach is broadly applicable to discovering efficient minimal gene sets for classifying cell types across a range of single-cell profiling applications.

Method

We developed a computational framework based on support vector machine (SVM) classifier to identify minimal gene sets that distinguish a set of cell-states in single-cell data. The procedure is an 'active' formulation of classical Sequential Feature Selection(SFS) [11]. In the conventional SFS approach, features are selected one-by-one in a greedy fashion to optimize an objective function. To reduce computational burden, we propose an active feature selection framework here, where we only use the subset of incorrectly classified cells at current step and then select the new feature based upon those cells. The active learning strategy enables efficient computation of minimal gene sets across large data sets by minimizing the total number of cells and genes that are analyzed.

In summary, our algorithm is applied to single-cell gene expression data and tasks, as input, gene expression data and cell-type labels, or alternately, we generate cell-type labels using unsupervised clustering. Our procedure, then, starts with an empty gene set, an empty cell set and a list of candidate genes and cells. The algorithm iteratively selects genes and classifies cells using identified genes by training a SVM model to classify cell-types. The algorithm identifies cells in the data set that classify poorly given the current gene set, and uses those cells to select additional genes to improve classification accuracy on the entire data set.

In single cell gene expression data, we use $x_i^{(j)} \in \mathbb{R}$ to denote the measurement of the j-th gene of the i-th cell and $x_i^{(D)} = [x_i^{(j)}]_{j \in D}$. Here, we assume the classification labels are given and consider a data-set $\{x_i, y_i\}_{i \in \{1,\ldots,N\}}$ contains N cells with total M genes, where $x_i = [x_i^{(j)}]_{j \in \{1,\ldots,M\}}$ and $y_i \in \mathbb{Z}^N$ are labels. The labels could be binary or multi-class. And we use J and I to refer to the set of selected genes and cells.

First, to seed the gene list, the algorithm selects c cells at random out of the total set of N cells and adds them to the cell-set. The parameter c is determined by the user. The algorithm, then, trains an SVM on the cell set, which defines an SVM margin w that optimally separates cells into classes that are consistent with labels on this seed set. A gene selection strategy we developed, max margin rotation, evaluates all candidate genes based on the margin w and one gene with the highest score is added to the gene set. A second SVM model is, then, learned given the current gene set, and, we identify cells that classify poorly given the current gene list. Then we sample c misclassified cells to identify genes that improve classification in the next step of our procedure. The integrated algorithm is shown in Algorithm 1.

The two novel components of the method are the strategies to evaluate and select genes and cells in each iteration. Specifically, we identify cells that classify poorly and use misclassified cells to identify highly informative genes. To select highly informative genes given the misclassified cells, a range of different strategies can be applied. In the conventional SVM, the procedure would sort features according to the absolute values of the components of weight w.[16] Here we use a new strategy we proposed, called Maximum Margin Rotation, which exhibits superior computational

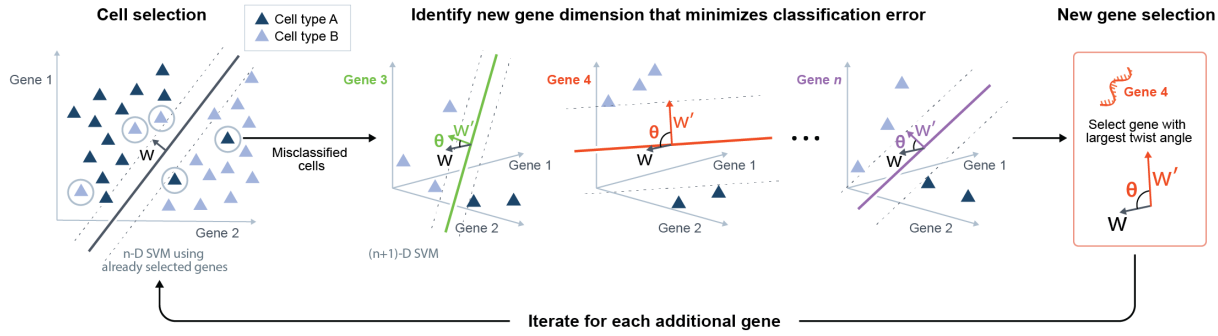


Figure 1: Description of Active SVM Feature Selection(ASFS). At the n-th step, an n-D SVM using all already selected genes is trained to select a certain number of misclassified cells, which is the cell selection step. In gene selection step, the least classifiable cells are taken as the training set. Based on this training set, N-n (n+1)-D SVMs are trained, where n dimensions are the same selected genes and the last dimension is one of the previously unselected candidate genes. Then we would obtain N-n weights w' corresponding N-n unselected genes as well as N-n margin rotation angle θ between every w' and the original weight w of the n-D SVM. The gene with the maximum rotation of margin is selected for the next round.

performance in biological datasets. Based on SVM, we can formalize the specific gene and cell selection strategies into two defined rules.

Assume the SVM classifier notation of one observation is $h_{w,b}(x_i^{(D)}) = g(w^T x_i^{(D)} + b)$ for any $i \in \{1, 2, \ldots, N\}$ and $D \subset \{1, 2, \ldots, M\}$ with respect to observation $x \in \mathbb{R}^{|D|}$, where $w \in \mathbb{R}^{|D|}$ and $b \in \mathbb{R}$ are parameters (the margin and bias respectively). Here, $g(z) = 1$ if $z \geq 0$, and $g(z) = -1$ otherwise. And the loss function is Hinge Loss[17] $\text{loss}_i = \max\{0, 1 - y_i h_{w,b}(x_i^{(D)})\}$, where $y_i \in \mathbb{R}$ is the ground truth label of observation x_i.

Cell selection: identification of maximally informative cells

For the cell selection strategy, we simply choose cells with largest SVM classification loss. We identify such cells through analysis of the dual form of the classical SVM classification problem. According to [18], after solving the primal optimization problem of soft margin SVM, we have the dual optimization problem with a non-negative Lagrange multipliers $\alpha_i \in \mathbb{R}$ for each inequality

constraint. Here $x_i^{(J)}$ refers to the measurement of the i-th cell with all selected genes.

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i_1,i_2=1}^{N} y_{i_1} y_{i_2} \alpha_{i_1} \alpha_{i_2} < x_{i_1}^{(J)}, x_{i_2}^{(J)} > \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{1}$$

where $C \in \mathbb{R}$ is a hyper-parameter we set to control the trade-offs between size of margin and margin violations when non-separable.

Then we solve the optimal solution α^* and apply the Karush-Kuhn-Tucker(KKT) dual-complementarity conditions[19] to obtain the following results where $w \in \mathbb{R}^{|J|}$ and the intercept term $b \in \mathbb{R}$ are optimal

$$\begin{aligned} \alpha_i^* = 0 \quad &\Rightarrow \quad y_i(w^T x_i^{(J)} + b) > 1 \\ \alpha_i^* = C \quad &\Rightarrow \quad y_i(w^T x_i^{(J)} + b) < 1 \\ 0 < \alpha_i^* < C \quad &\Rightarrow \quad y_i(w^T x_i^{(J)} + b) = 1. \end{aligned} \tag{2}$$

Therefore, for each cell, the Lagrange multipliers α_i indicates whether the cell falls within the SVM margin defined by the vector w. $\alpha_i > 0$ means $y_i(w^T x_i + b) \leq 1$, i.e. cells are on or inside SVM margin. Hence, we can directly select cells with $\alpha_i > 0$. In practice, normally we only select cells with $\alpha_i = C$, which means incorrectly classified cells.

Using this mathematical formulation, we were able to develop two different versions of the AFSF procedure, min-complexity strategy and min-cell strategy, for different purposes. The min-complexity strategy minimizes the computational complexity of the framework. A certain fixed number of cells is randomly selected among all misclassified cells each iteration. Therefore, a small number of cells can be analyzed at each round and typically few cells are used repeatedly. In the min-cell strategy, the procedure minimizes the total number of unique cells we acquire during the procedure. Therefore, the misclassified cells already used in previous steps are selected with first priority. In details, assume we should use c cells for each iteration and there are $a + b$ misclassified cells at current iteration, where a cells are used at least one time in previous iterations while b cells are newly selected. If $a \geq c$, we should randomly select c cells among these a cells we used before. If $a < c$, we should add these a used cells and randomly select $c - a$ cells in b new cells. When using min-cell strategy, cells tend to be re-used many times and the curve of number of unique cells we acquired should be convergent along with the number of genes we select. In experiments, the number of cells for each step, c, is a hyper-parameter set by the user. Typically, the parameter can be set to a small number using min-complexity strategy as a sufficient number of new cells is

considered in the procedure. Selecting a small number of cells each round reduces computational complexity. In the min-cell strategy it can be advantageous to select a larger number of total cells to guarantee diversity of training cells.

Another consideration is to balance the number of cells sampled of classes to avoid rapidly diminishing results for small clusters. Therefore, besides randomly sampling misclassified cells of every classes, we have another cell sampling strategy called 'balanced' sampling. Specifically, assume there are Z classes and we should use c cells for each iteration. Assume there are $a = \sum_{z=1}^{Z} a_z$ misclassified cells at current iteration, where $a_1 < a_2 < \ldots < a_Z$ are the number of cells in Z classes. If $a_1 \geq c/Z$, we should randomly sample c/Z cells for each cell type. Otherwise, if $a_1 < c/Z$, we should add all a_1 cells. Then if $a_2 < (c - a_1)/(Z - 1)$, we should add all a_2 cells. This procedure repeats until encountering a class, for example class z_0, with sufficient misclassified cells. Then $(c - \sum_{z=1}^{z_0-1} a_z)/(Z - z_0 + 1)$ misclassified cells should be sampled at random from each rest class.

Gene selection by maximizing margin rotation

To select genes, we analyze the misclassified cells to identify genes that will induce the largest rotation of the margin, which inspired by the popular active learning method, Expected Model Change[20]. We quantify rotation of the margin by calculating the twist angle induced in w when we add a new dimension (gene) to the classifier. Assume J is the set of genes we have selected so far. Once we add a gene into the $|J|$-dimensional data space, the parameter w will have one more dimension. The rotation of margin measures how much w twists after adding the new dimension compared with weight in the previous iteration.

Specifically, assume J is the set of genes we have selected so far. According to [18], we derive the corresponding w from the optimal solution α^*. After solving the dual optimization problem (1), we have:

$$w = \sum_{i=1}^{N} \alpha_i^* y_i x_i^{(J)}. \tag{3}$$

Then we pad w with zero to get a $|J + 1|$-dimensional weight w_{padded}, whose first $|J|$ dimensions is w and the $|J + 1|$-th dimension is zero.

For each candidate gene j, we train a new $|J + 1|$-dimensional SVM model and have weight w_j,where $j \in \{1, \ldots, M\} \setminus J$. That is to say, for candidate gene j, we solve the dual optimization problem (4) and find a new optimal multiplier $\alpha^{*(j)}$. Note that we only use the selected cells here,

$i_1, i_2 \in I$. Then we have w_j as shown in equation (5)

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i \in I} \alpha_i^{(j)} - \frac{1}{2} \sum_{i_1, i_2 \in I} y_{i_1} y_{i_2} \alpha_{i_1}^{(j)} \alpha_{i_2}^{(j)} \langle x_{i_1}^{(J \cup \{j\})}, x_{i_2}^{(J \cup \{j\})} \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i^{(j)} \leq C \\ & \sum_{i \in I} \alpha_i^{(j)} y_i = 0 \end{aligned} \tag{4}$$

$$w_j = \sum_{i \in I} \alpha_i^{*(j)} y_i x_i^{(J \cup \{j\})} \tag{5}$$

The angle θ_j between w_j and w_{padded} is the expected angle the margin rotates with corresponding to the j-th candidate gene. Then the j-th gene with largest angle θ_j will be selected. We measure the angle between two vectors using cosine similarity[21]:

$$\vartheta_j = \arccos \cos \vartheta_j = \arccos \frac{\langle w_j, w_{padded} \rangle}{\| w_j \| \| w_{padded} \|} \tag{6}$$

Therefore, a new gene which maximizes ϑ_j is selected to maximize the expected model change.

For multi-class classification, the SVM is handled according to a one-vs-rest scheme[22], where a separate classifier is fit for each class, against all other classes. Margin rotation is represented as the sum of weight components in each class dimension. Hence with Z classes, we get Z weight components corresponding to Z one-vs-the-rest classification decision boundaries. Assume the weight component for class z of previous $|J|$-dimensional SVM model is $w^{(z)}$. Denote the $|J+1|$-dimensional weight after zero-padding of $w^{(z)}$ as $w_{padded}^{(z)}$ and the new $|J+1|$-dimensional weight component of class z with j-th gene as $w_j^{(z)}$, where $z \in 1, \ldots, Z$. Then we have:

$$\vartheta_j^{(z)} = \arccos \cos \vartheta_j^{(z)} = \arccos \frac{\langle w_j^{(z)}, w_{padded}^{(z)} \rangle}{\| w_j^{(z)} \| \| w_{padded}^{(z)} \|} \tag{7}$$

$$\vartheta_j = \sum_{z=1}^{Z} \vartheta_j^{(z)} \tag{8}$$

Computational complexity

The computational complexity of the complete procedure mainly depends on the training of SVM. The standard computational complexity of SVM can be as small as $\mathcal{O}(MN^2)$.[23] Assume that we plan to select $k \in \mathbb{N}$ genes in total and use $c \in \mathbb{N}$ poorly classified cells at each step, where $k \ll M$ and $c \ll N$ are constants. The computational complexity of ASFS should be

$\mathcal{O}(\sum_{i=1}^{k}(i \cdot N^2 + (M+1-i) \cdot i \cdot c^2)) \sim \mathcal{O}(N^2 + M)$. But in practice, we implement ASFS using the linear SVM library LIBLINEAR[24], whose computational complexity is $\mathcal{O}(MN)$. Therefore, the corresponding complexity of ASFS is $\mathcal{O}(M+N)$. This is a significant improvement in marker gene selection methods.

Algorithm 1: Active Linear SVM Gene Selection

Input: $c, k \in \mathbb{N}$, $J = \emptyset$
Output: J
Randomly select c cells $I \subset \{1, \ldots, N\}, |I| = c$
Train a 1-D SVM model on training set I for each candidate gene: $\{h_{w,b}^{(j)}\}_{j \in \{1,\ldots,M\}}$
$loss_j = \sum_{i \in I} max\{0, 1 - y_i h_{w,b}^{(j)}(x_i^{(j)})\}$
Select one gene $j_0 \in \{1, \ldots, M\}$ with lowest $loss_j$
$J = J \cup \{j_0\}$
repeat
  Optimize (1) and get optimal solution $\{\alpha_i^*\}_{i=1}^N$
  Randomly select c cells $I \subset \{1, \ldots, N\}$ with $\alpha_i^* = C$, where $|I| = c$
  $w = \sum_{i=1}^N \alpha_i^* y_i x_i^{(J)}$
  $w_{padded} = [w, 0]$
  For each $j \in \{1, \ldots, M\} \setminus J$, optimize (4) and get optimal solution $\{\alpha_i^{*(j)}\}_{i \in I}$
  $w_j = \sum_{i \in I} \alpha_i^{*(j)} y_i x_i^{(J \cup \{j\})}$
  $\vartheta_j = \arccos \cos \vartheta_j = \arccos \frac{<w_j, w_{padded}>}{\|w_j\| \|w_{padded}\|}$
  Select one gene $j^* \in \{1, \ldots, M\} \setminus J$ with largest ϑ_j
  $J = J \cup \{j^*\}$
until $|J| \geq k$

Results

We test ASFS method on three biology datasets: a dataset of peripheral blood mononuclear cells (PBMCs) [13], the Tabula Muris dataset[14], and MM datasets [15]. We test with both min-cell strategy and min-complexity strategy introduced in Cell Selection part. The figures of Min-complexity strategy contain both results that balance the number of cells for every classes or not. The results for both strategies contain four parts: 1) We show the classification accuracy of test set along with the number of genes we select and compare the performance to several widely-used feature selection methods, including passive SVM, conventional SVM, correlation coefficient[25], mutual information[26], Chi-square[27], feature importance by decision tree[28], and randomly sample genes, showing that ASFS performs best. All of the comparison methods select genes one by one. At each iteration, they select a new gene with largest score in terms of corresponding evaluation functions and use the same number of cells as our method. But they randomly sample cells at each iteration without active learning approaches. Here the method called passive SVM

chooses genes with the max margin rotation strategy we proposed. Conventional SVM also called naive SVM selects the gene with largest weight component, which is the most popular SVM feature selection method. 2) In each case, we provide the plots of total number of unique cells we used vs. the number of genes during the procedure, which helps to understand the differences and advantages of min-cell and min-complexity strategies. 3) To indicate the cell type clusters extracted with the gene set ASFS got, We added the t-SNE projection computed using genes we select as well as using the entire original filtered gene expression matrix for comparison. 4) We choose 40 genes that provide interesting biological information in these tasks and show their t-SNE plots highlighting gene expression values using original t-SNE coordinates.

Active feature selection on human PBMC data

To test the performance of the method, we applied our active feature selection method to extract minimal classifying gene sets for human PBMCs. Specifically, we applied the method to a data set from [13] that contains 10194 cells profiled via single-cell transcriptional profiling. We used Louvain clustering[29] to identify T-cells, activated T/NK cells, B-cells, and Monocytes (Figure 2A).

Our active SVM feature selection(ASFS) strategy iteratively selects cells and constructs a gene list while calculating the classification accuracy on the multi-task SVM classification task. On the PBMC data, the procedure constructs gene sets of increasing size that can reproduce cell type clusters (Figure 2B) and classify the 5 major cell-types at greater than 85% accuracy with as few as 15 genes (Figure 2C). On this classification task, aSVM outperforms than other methods including conventional Sequential Feature Selection, conventional SVM, correlation coefficient [25], mutual information[26], Chi-square[27], and feature importance by decision tree[28].

A key benefit of the active learning strategy is that a relatively small fraction of the data set is analyzed, so that the procedure can generate these gene sets while only computing across 200-300 cells (Figure 2D). At each iteration, a set number of misclassified cells ($n = 100$) are selected but the total number of cells used does not increase in increments of 100, since some cells are repeatedly misclassified and are thus repeatedly used for each iteration.

In addition to enabling classification of the cell-types in the data set, the ASFS gene sets provide a low-dimensional space in which to analyze the data. When we reduced our analysis to consider only the top 100 genes selected by the ASFS algorithm, we can generate a low-dimensional representations of the cell population (t-SNE) that preserve important structural features of the data including the distinct cell-type clusters [Figure 2B].

The procedure generates gene sets that contain known and novel markers, each plotted individually in a t-SNE grid [Figure 2E]. For instance, MS4A1 and CD79 are well-established B-cell markers, and IL7R and CD3G are well-established T-cell markers. However, we also find interesting genes which are not commonly used as markers, but whose expression is cell-type specific. For instance,

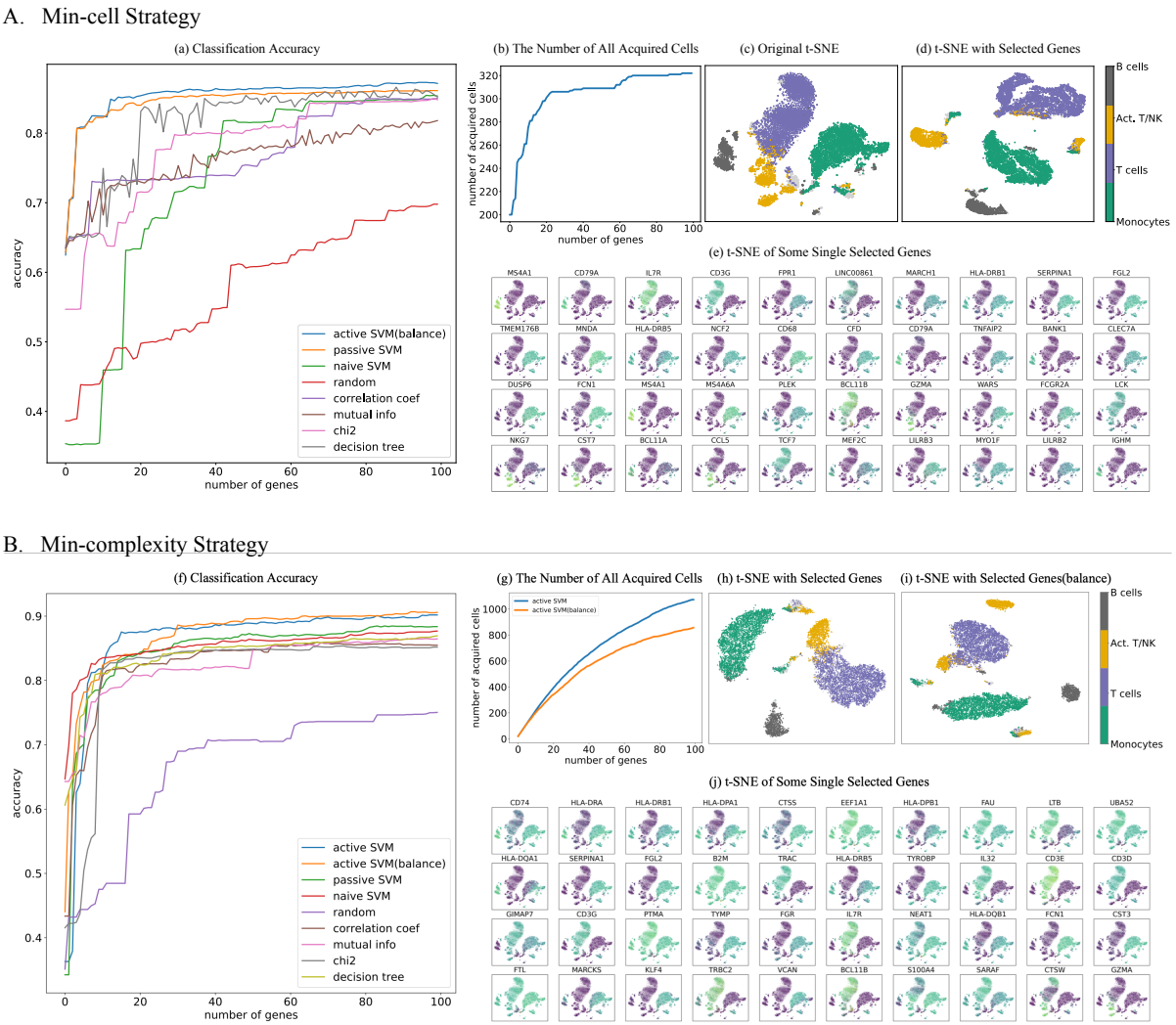


Figure 2: Gene selection experiment results of cell type classification on PBMC dataset. Figure A shows the results up to 100 selected genes using Min-cell strategy with 100 cells are selected at each iteration. Figure B shows 100 selected genes using Min-complexity strategy using 20 cells each iteration. The subplots contain: classification accuracy vs. gene set size using min-cell strategy (a) and min-complexity strategy (f); the total number of unique cells used vs. gene set size with min-cell strategy (b) and min-complexity strategy (g); the t-SNE plots of entire filtered dataset (c), gene set selected by Min-cell strategy (d), and gene set selected by Min-complexity strategy with randomly sampling (h) and 'balanced' sampling (i); t-SNE of gene expression values for 40 interesting genes selected with min-cell strategy (e) and min-complexity strategy (j).

we find highly monocyte-specific expression of FPR1, which encodes N-formylpeptide receptor, which was recently discovered to be the receptor for plague effector proteins [30]. We also find T-cell/NK-cell specific expression of a long noncoding RNA, LINC00861, whose function is unknown but has been correlated with better patient outcome in lung adenocarcinoma [31]. Identified marker genes generally provide high-specificity for individual cell types, but some mark multiple cell types (i.e. MARCH1, which marks Monocytes and B-cells).

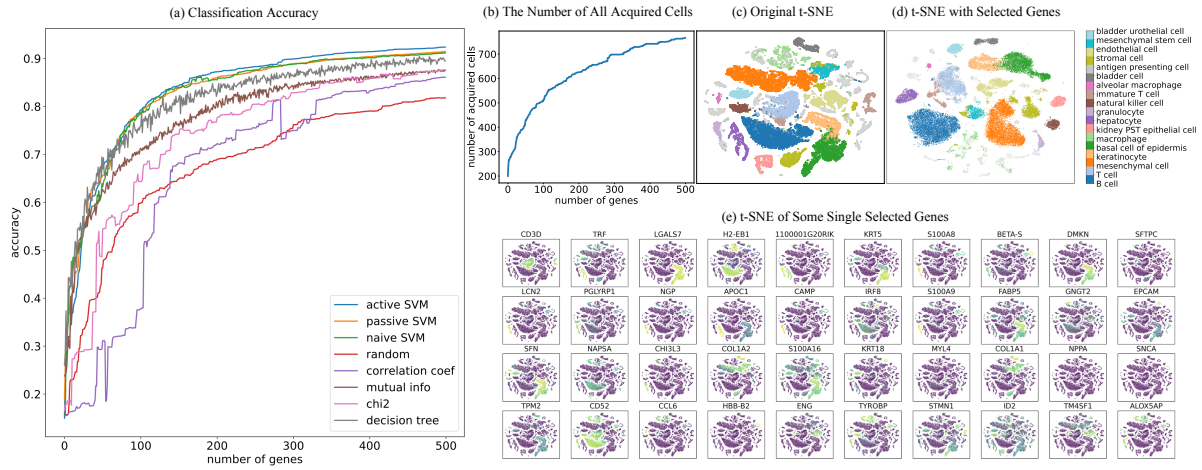
Identifying optimal gene sets for cell-type classification in Tabula Muris tissue survey

To test the performance of our active SVM feature selection(ASFS) method at larger-scale, we applied it to a dataset that includes a large number of unique cell types: the Tabula Muris mouse tissue survey [14]. We used droplet-based scRNA-sequencing data from Tabula Muris, which contains 55,656 single cells across 58 annotated cell types, and 12 major tissues. In our assessment, we used the supplied cell type labels, agnostic of tissue type. Thus, cells labeled 'macrophage' from the Spleen are considered as belonging to the same class as cells labeled 'macrophage' from the Mammary Gland.

Even with a large number of cell types, our active SVM feature selection(ASFS) method can construct gene sets that achieve high accuracy ($> 90\%$), compared to other methods (Fig 3a). To construct a gene set of size n=500, ASFS used fewer than 800 unique cells (Fig 3b), which works out to an average of ~ 14 cells/cell type. Just as with the PBMC data, we show we recreate the clustering patterns from the original data (Fig 3c) when analyzing the cells within the low-dimensional t-SNE space spanned by the selected 500 genes (Fig 3d).

Our approach allows us to construct an informative set of marker genes for identifying mouse cell types across disparate tissues. Even when computing across a large number of cell types, we can identify highly cell type specific genes, such as CD3D, a well-established T-cell marker, or TRF (transferrin), which is known to be selectively secreted by hepatocytes[32], or LGALS7 (galectin-7), which is known to be specific for basal and differentiated cells of stratified epithelium [33]. However, given the functional overlap between different cell types, the informative genes within our set include many that mark multiple cell types. For instance, we see that H2-EB1[34], a protein important in antigen presentation, expressed in both B-cells and Macrophages, both of which are considered professional antigen presenting cells (APCs). Our analysis also identifies new cell type-specificity for relatively unknown genes, such as granulocyte-, and hepatocyte- specific expression of 1100001G20RIK (also known as Wdnm-like adipokine), which had previously only been associated with adipocytes [35].

A. Min-cell Strategy



B. Min-complexity Strategy

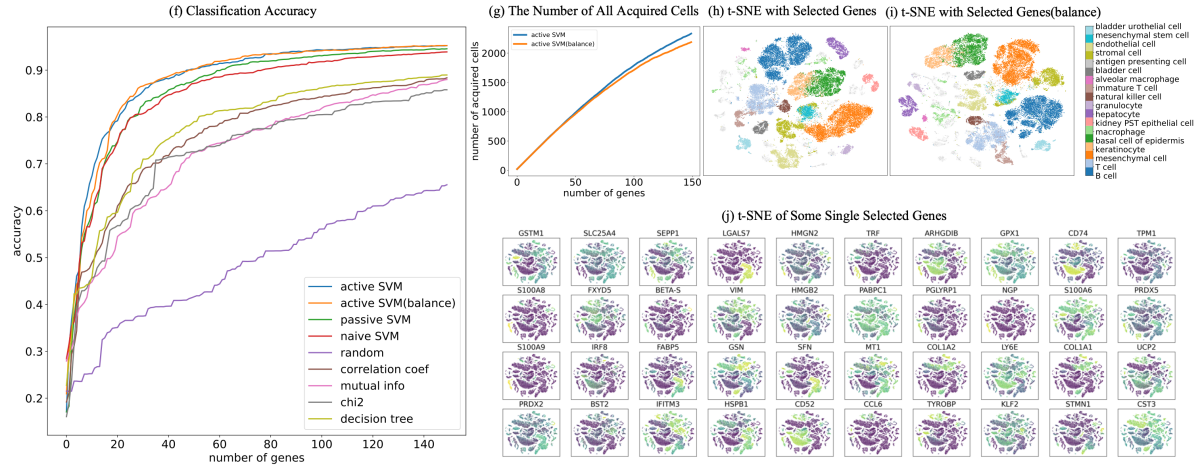


Figure 3: Identification of minimal gene set for cell type classification in Tabula Muris mouse tissue survey Figure A shows classification results of 500 genes selected using Min-cell strategy with 200 cells per iteration. Figure B shows 150 genes selected by min-complexity strategy using 20 cells each iteration. The results are shown as Figure 2 in accuracy plot, t-SNE, etc.

Minimal gene sets for classification of disease state in peripheral blood cells from multiple myeloma patient samples

Having demonstrated the method on identifying cell-type specific markers at both small- and large-scale, we next turned to applying ASFS to discovering disease-specific markers. We used single-cell data from peripheral blood immune cells collected from both healthy donors and patients who have been diagnosed with multiple myeloma [15]. Multiple myeloma (MM) is an incurable cancer of plasma cells, known as myeloma cells, that overproliferate in the bone marrow. Although the myeloma cells are typically the target of analysis because they are the causative agent of disease, the peripherally circulating immune cells also contain signatures of disease. Known signatures in the peripheral blood include a depleted B-cell population [36, 37], an increased myeloid-derived suppressor cell count [38], as well as T-cell immunosenescence [39, 37].

In our analysis, we sought to further define transcriptional markers that distinguish healthy peripheral immune cells from the cells of MM patients. This dataset contains immune cells from two healthy donors as well as four MM donors. We performed feature selection using heterogeneous populations of cells labeled only by disease state. We compared the classification accuracy for the ASFS method versus other methods (Fig 4a). We found that the ASFS reached high accuracy in a limited number of steps and consistently outperformed other methods. We also ran the ASFS method with two cell sampling strategies, randomly sampling, and 'balanced' sampling, in which equal numbers of cells from each cell type are sampled to correct for artifacts due to different cell type proportions between samples. We noted that although the balanced approach (orange line) gave higher classification accuracy at early iterations, these differences are no longer apparent after selecting 20 genes.

A t-SNE projection of the original dataset shows completely non-overlapping clusters between healthy and MM cells (Fig 4b). These non-overlapping clusters are replicated in t-SNEs constructed from forty genes selected using both the random sampling strategy (Fig 4c) as well as the 'balanced' cell sampling strategy (Fig 4d).

We analyzed the function of the genes identified by the ASFS method and saw that most regulate housekeeping functions, suggesting that global shifts in translation and motility are disrupted in multiple myeloma patients. Translation associated markers include Eukaryotic Translation Initiation Factor 1 (EIF1), Eukaryotic Translation Elongation Factor 1 Alpha 1 (EEF1A1), and prefoldin subunit 5 (PFDN5). Motility associated genes include ACTB, putative anti-adhesion molecule CD52, and actin-sequestering protein TMSB4X.

We also found both known and potentially novel markers of MM within the peripheral blood immune cells. Our analysis identifies TPT1, which has been previously associated with MM [40]. We also find that RACK1 (also known as GNB2L1), a scaffolding protein that coordinates critical functions including cell motility, survival and death, is broadly upregulated in peripheral immune cells from MM patients. Although this gene has been previously associated with myeloma cells [41], its regulation had not been previously reported in peripherally circulating immune cells. Our

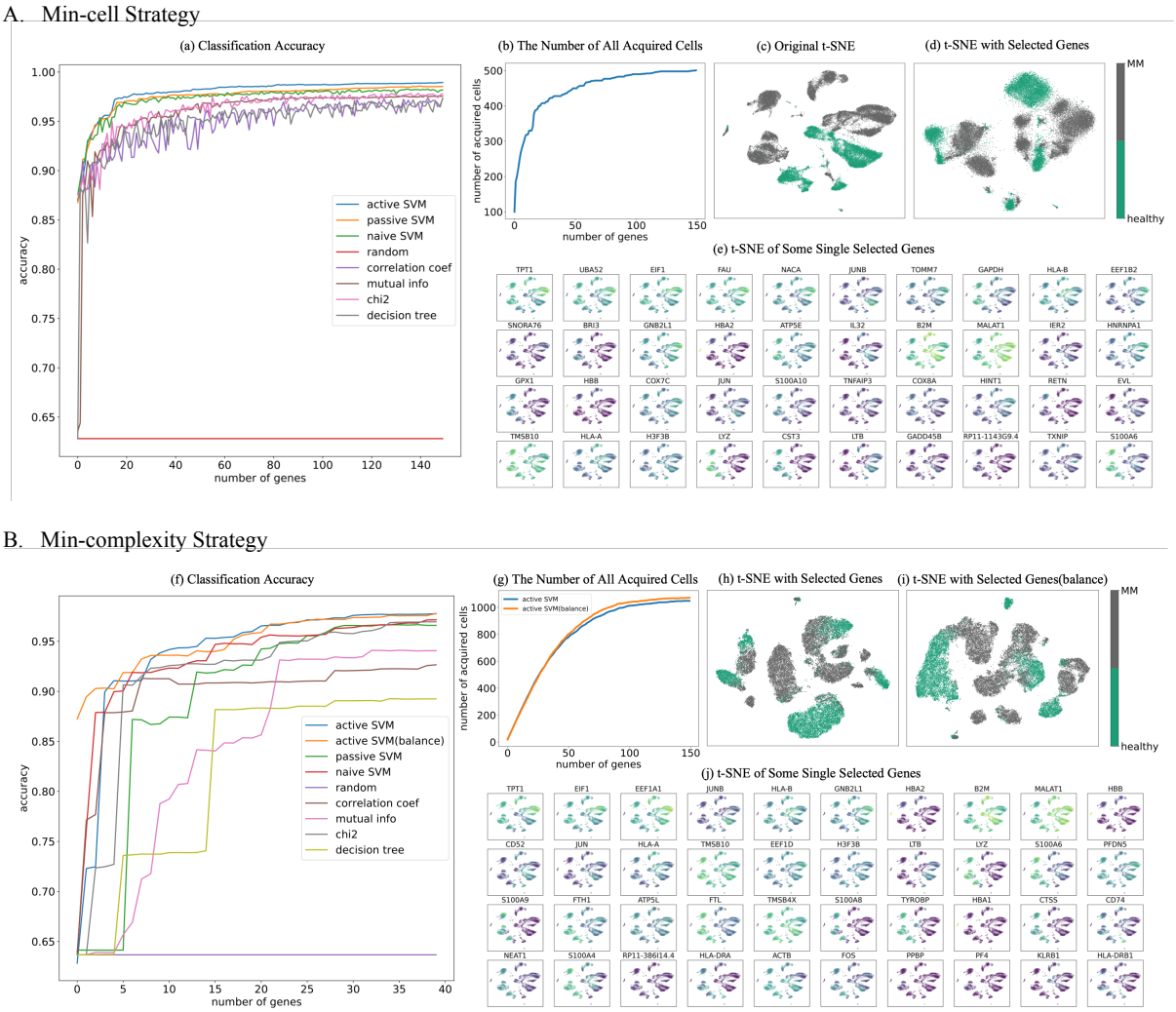


Figure 4: Gene selection experiment results of healthy-disease classification on MM dataset. Figure A shows classification results of 150 genes selected using Min-cell strategy with 100 cells per iteration. Figure B shows only 40 genes selected by min-complexity strategy using 20 cells each iteration. The results are shown as Figure 2 in accuracy plot, t-SNE, etc.

ability to discover MM-specific genes within peripheral immune cells suggests a broader utility for discovering disease-specific genes across many different types of pathologies.

Discussion

In this paper, we have introduced an active feature selection procedure for discovery of minimal gene sets using single-cell mRNA-seq data. Our procedure was designed to identify minimal gene sets in large data sets. Our procedure minimizes the total number of cells and genes that must be analyzed at each iteration. We demonstrate that ASFS is able to identify small gene sets with tends to hundreds of genes that still enable highly accurate cell-type classification. In future work, we aim to demonstrate that the gene sets can be applied to create experimental gene targeting panels. The core components of our framework can be extended to methods other than SVM as well as to develop strategies to improve the on-line acquisition of single-cell data.

Data Availability

The PBMC Single-cell RNA-seq data have been deposited in the Short Read Archive under accession number SRP073767 by the authors of [13]. Data are also available at `http://support.10xgenomics.com/single-cell/datasets`.

The original Tabula Muris dataset is available at `https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733`.

The original multiple myeloma PBMC dataset, containing 2 healthy donors and 4 multiple myeloma donors, is available at `https://figshare.com/articles/dataset/PopAlign_Data/11837097/3`.

Code Availability

The example source codes including algorithm and preprocessing codes are publicly available on GitHub at (`https://github.com/xqchen/Active-feature-selection-in-single-cell-mR`

References

[1] H. C. Fan, G. K. Fu, and S. P. Fodor, "Combinatorial labeling of single cells for gene expression cytometry," *Science*, vol. 347, no. 6222, 2015.

[2] J. M. Replogle, T. M. Norman, A. Xu, J. A. Hussmann, J. Chen, J. Z. Cogan, E. J. Meer, J. M. Terry, D. P. Riordan, N. Srinivas, *et al.*, "Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing," *Nature biotechnology*, pp. 1–8, 2020.

[3] J. L. Marshall, B. R. Doughty, V. Subramanian, P. Guckelberger, Q. Wang, L. M. Chen, S. G. Rodriques, K. Zhang, C. P. Fulco, J. Nasser, *et al.*, "Hypr-seq: Single-cell quantification of chosen rnas via hybridization and sequencing of dna probes," *Proceedings of the National Academy of Sciences*, vol. 117, no. 52, pp. 33404–33413, 2020.

[4] G. M. Sheynkman, K. S. Tuttle, F. Laval, E. Tseng, J. G. Underwood, L. Yu, D. Dong, M. L. Smith, R. Sebra, L. Willems, *et al.*, "Orf capture-seq as a versatile method for targeted identification of full-length isoforms," *Nature communications*, vol. 11, no. 1, pp. 1–12, 2020.

[5] K. A. Riemondy, M. Ransom, C. Alderman, A. E. Gillen, R. Fu, J. Finlay-Schultz, G. D. Kirkpatrick, J. Di Paola, P. Kabos, C. A. Sartorius, *et al.*, "Recovery and analysis of transcriptome subsets from pooled single-cell rna-seq libraries," *Nucleic acids research*, vol. 47, no. 4, pp. e20–e20, 2019.

[6] C. Delaney, A. Schnell, L. V. Cammarata, A. Yao-Smith, A. Regev, V. K. Kuchroo, and M. Singer, "Combinatorial prediction of marker panels from single-cell transcriptomic data," *Mol. Syst. Biol.*, vol. 15, p. e9005, Oct. 2019.

[7] F. Wang, S. Liang, T. Kumar, N. Navin, and K. Chen, "Scmarker: ab initio marker selection for single cell transcriptome profiling," *PLoS computational biology*, vol. 15, no. 10, p. e1007445, 2019.

[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] R. M. Felder and R. Brent, "Active learning: An introduction," *ASQ higher education brief*, vol. 2, no. 4, pp. 1–5, 2009.

[11] T. Rückstieß, C. Osendorfer, and P. van der Smagt, "Sequential feature selection for classification," in *Australasian Joint Conference on Artificial Intelligence*, pp. 132–141, Springer, 2011.

[12] W. S. Noble, "What is a support vector machine?," *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[13] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature communications*, vol. 8, no. 1, pp. 1–12, 2017.

[14] T. M. Consortium *et al.*, "Single-cell transcriptomics of 20 mouse organs creates a tabula muris.," *Nature*, vol. 562, no. 7727, p. 367, 2018.

[15] S. Chen, P. Rivaud, J. H. Park, T. Tsou, E. Charles, J. R. Haliburton, F. Pichiorri, and M. Thomson, "Dissecting heterogeneous cell populations across drug and disease conditions with popalign," *Proceedings of the National Academy of Sciences*, vol. 117, no. 46, pp. 28784–28794, 2020.

[16] Y.-W. Chang and C.-J. Lin, "Feature ranking using linear svm," in *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008* (I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, eds.), vol. 3 of *Proceedings of Machine Learning Research*, (Hong Kong), pp. 53–64, PMLR, 03–04 Jun 2008.

[17] Wikipedia contributors, "Hinge loss — Wikipedia, the free encyclopedia," 2021. [Online; accessed 11-March-2021].

[18] L. Bottou and C.-J. Lin, "Support vector machine solvers," *Large scale kernel machines*, vol. 3, no. 1, pp. 301–320, 2007.

[19] G. Gordon and R. Tibshirani, "Karush-kuhn-tucker conditions," *Optimization*, vol. 10, no. 725/36, p. 725, 2012.

[20] B. Settles, "Active learning literature survey," 2009.

[21] P. Xia, L. Zhang, and F. Li, "Learning similarity with cosine similarity ensemble," *Information Sciences*, vol. 307, pp. 39–52, 2015.

[22] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[23] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *the Journal of machine Learning research*, vol. 9, pp. 1871–1874, 2008.

[25] R. Taylor, "Interpretation of the correlation coefficient: a basic review," *Journal of diagnostic medical sonography*, vol. 6, no. 1, pp. 35–39, 1990.

[26] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[27] M. L. McHugh, "The chi-square test of independence," *Biochemia medica: Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[28] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[29] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[30] P. Osei-Owusu, T. M. Charlton, H. K. Kim, D. Missiakas, and O. Schneewind, "FPR1 is the plague receptor on host immune cells," *Nature*, vol. 574, pp. 57–62, Oct. 2019.

[31] A. P. Sage, K. W. Ng, E. A. Marshall, G. L. Stewart, B. C. Minatel, K. S. S. Enfield, S. D. Martin, C. J. Brown, N. Abraham, and W. L. Lam, "Assessment of long non-coding RNA expression reveals novel mediators of the lung tumour immune response," *Sci. Rep.*, vol. 10, p. 16945, Oct. 2020.

[32] W. Guan, Z. Gao, C. Huang, M. Fang, H. Feng, S. Chen, M. Wang, J. Zhou, S. Hong, and C. Gao, "The diagnostic value of serum DSA-TRF in hepatocellular carcinoma," *Glycoconj. J.*, vol. 37, pp. 231–240, Apr. 2020.

[33] T. Magnaldo, D. Fowlis, and M. Darmon, "Galectin-7, a marker of all types of stratified epithelia," *Differentiation*, vol. 63, pp. 159–168, July 1998.

[34] M. J. Stables, S. Shah, E. B. Camon, R. C. Lovering, J. Newson, J. Bystrom, S. Farrow, and D. W. Gilroy, "Transcriptomic analyses of murine resolution-phase macrophages," *Blood*, vol. 118, pp. e192–208, Dec. 2011.

[35] Y. Wu and C. M. Smas, "Wdnm1-like, a new adipokine with a role in MMP-2 activation," *Am. J. Physiol. Endocrinol. Metab.*, vol. 295, pp. E205–15, July 2008.

[36] A. C. Rawstron, F. E. Davies, R. G. Owen, A. English, G. Pratt, J. A. Child, A. S. Jack, and G. J. Morgan, "B-lymphocyte suppression in multiple myeloma is a reversible phenomenon specific to normal b-cell progenitors and plasma cell precursors," *British journal of haematology*, vol. 100, no. 1, pp. 176–183, 1998.

[37] R. J. P. de Magalhães, M.-B. Vidriales, B. Paiva, C. Fernandez-Gimenez, R. García-Sanz, M.-V. Mateos, N. C. Gutierrez, Q. Lecrevisse, J. F. Blanco, J. Hernández, *et al.*, "Analysis of the immune system of multiple myeloma patients achieving long-term disease control by multidimensional flow cytometry," *Haematologica*, vol. 98, no. 1, p. 79, 2013.

[38] E. Malek, M. de Lima, J. J. Letterio, B.-G. Kim, J. H. Finke, J. J. Driscoll, and S. A. Giralt, "Myeloid-derived suppressor cells: The green light for myeloma immune escape," *Blood reviews*, vol. 30, no. 5, pp. 341–348, 2016.

[39] H. Suen, R. Brown, S. Yang, C. Weatherburn, P. J. Ho, N. Woodland, N. Nassif, P. Barbaro, C. Bryant, D. Hart, *et al.*, "Multiple myeloma causes clonal t-cell immunosenescence: identification of potential novel targets for promoting tumour immunity and implications for checkpoint blockade," *Leukemia*, vol. 30, no. 8, pp. 1716–1724, 2016.

[40] F. Ge, L. Zhang, S.-C. Tao, K. Kitazato, Z.-P. Zhang, X.-E. Zhang, and L.-J. Bi, "Quantitative proteomic analysis of tumor reversion in multiple myeloma cells," *Journal of proteome research*, vol. 10, no. 2, pp. 845–855, 2011.

[41] T. Xiao, W. Zhu, W. Huang, S.-S. Lu, X.-H. Li, Z.-Q. Xiao, and H. Yi, "Rack1 promotes tumorigenicity of colon cancer by inducing cell autophagy," *Cell death & disease*, vol. 9, no. 12, pp. 1–13, 2018.