Single Cell Characterization of Multiple Myeloma Driver Genes Using a Machine Learning Approach

By Brandon Kim and Helen Tang

William A. Shine Great Neck South High School

Abstract: Multiple myeloma (MM) is a clonal cell cancer characterized by excessive cell division of plasma cells in the bone marrow, which can then overcrowd healthy cells. As a result, end organ damage to kidneys, bones, and the liver occurs. The worldwide incidence of MM amounted to 160,000 cases in 2018 and 106,000 patients have succumbed to the disease. MM is diagnosed relatively well by detecting M monoclonal protein produced from cancerous cells, yet mortality rates remain high because there is a lack of a specific treatment. By identifying upregulated genes found in malignant plasma cells, scientists can develop new and stronger therapies tailored to potential driver genes. This study takes a novel machine learning approach to identify driver genes of MM. A single-cell RNA sequencing dataset obtained from Gene Expression Omnibus containing data from 29,367 plasma cells and 22,088 genes was utilized in this study. This study evaluated the performance of three machine learning models: Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), with RF achieving the highest accuracy of 95.61%.To name a few genes, the models identified ANKRD28 and HLA-DPA1 as potential driver genes that have been cross-validated with previous literature. Notably, the models identified RP5-1171I10.5–a gene not yet established to be associated with multiple myeloma which shows potential to be further studied for research. These genes show potential to be further studied for specific targeted genetic therapy.

1. Introduction

1.1 Multiple Myeloma

Multiple Myeloma (MM) is a clonal plasma cell cancer characterized by excessive cell proliferation and abnormal antibody formation. Overproduction of abnormal proteins crowds out healthy cells and results in numerous health complications. Unchecked, MM ultimately leads to end organ damage, renal dysfunction, hypercalcemia, bone disease, and peripheral neuropathy (Albagoush et al., 2023). In the United States alone, MM prevalence continues to rise each year, and it is estimated that 35,730 new cases and 12,590 deaths are expected to occur in 2023 (American Cancer Society). Despite the growing concern of MM deaths, survival rates continue to remain low with only 59.8% of patients surviving within 5 years (American Cancer Society).

Researchers have indicated that risk factors such as obesity, chronic inflammation, and radiation exposure increase the likelihood of developing the disease. On the molecular level, MM is caused due to numerous genetic mutations, epigenetic modification, and abnormal miRNA, however the chief cause remains unknown (Das et al., 2022). Genetic changes such as trisomies and translocations of the immunoglobulin heavy chain locus on chromosome 14 are primary abnormalities that occur in the first stage of MM progression. Changes in genetic composition result in the formation of abnormal antibodies, such as the Myeloma monoclonal protein (M protein) (Kyle et al., 2011). M protein is mainly responsible for overcrowding healthy cells which results in organ damage. As a result, researchers have established the rise of M monoclonal protein levels in the blood as the major indicator for MM progression. However, the exact genes responsible for abnormal M protein production remains unknown.

Multiple myeloma progresses in four distinct stages. First, Normal Bone Marrow cells (NBM) progress into Monoclonal Gammopathy of Undetermined Significance (MGUS). At this

stage, M protein levels have not yet been detected in the blood. Instead, general serum monoclonal protein level produced from abnormal plasma cells must reach a level above 3 gm/dL in the blood. Additionally, clonal bone marrow plasma cells below 10% of all plasma cells and absence of end organ damage must be identified to qualify a patient to be in the MGUS stage (Rajkumar, 2022). This stage is considered benign and health complications have not yet occurred. There are no distinct biomarkers found in this stage for disease prognosis except for increased levels of serum blood protein. Nonetheless, MGUS still displays clinical importance since roughly 20% of all MGUS patients develop MM later on in life (Mateos et al., 2020).

MGUS cells eventually progress to the next stage, Smoldering Multiple Myeloma (SMM). This stage is characterized by even higher levels of serum monoclonal protein and is distinguished from MGUS by the now present M protein in the blood. To qualify for SMM, M protein levels in the blood must exceed 3 gm/dL (Rajkumar, 2022). Concern begins at this stage as roughly 10% of SMM patients develop MM within the first five years (Mateos et al., 2020).

SMM finally develops into the active Multiple Myeloma, the last stage of MM progression. Clonal bone marrow plasma cells exceed 10% of all plasma cells and evidence of end organ damage is identified in this stage (Michels et al., 2017).

1.2 Diagnosis and Treatment of Multiple Myeloma

When a patient is suspected to have a presence of M protein, a combination of tests are run to measure serum blood protein levels. Tests such as the serum protein electrophoresis, serum immunofixation, and serum FLC assay all measure M protein levels to identify if a patient is in one of the stages of MM progression (Rajkumar, 2022). Despite diagnosis of MM being relatively advanced, researchers have still yet to develop an advanced targeted therapy for MM.

The main and most common treatment for MM is the use of proteasome inhibitors (Multiple Myeloma Research Foundation). Common proteasome inhibitors such as bortezomib, carfilzomib, and ixazomib are responsible for preventing proteasomes in the cell from breaking down pro-apoptotic factors (Sharma and Preuss, 2022). Specifically, bortezomib reversibly binds to the chymotrypsin subunit section of the 26S proteasome. Eventually, the buildup of pro-apoptotic factors activates programmed cell death pathways in cancerous cells.

Although these treatments work to some degree in killing MM cells, they act generally and are inefficient. Since they do not act to target specific metabolic pathways or genes characteristic of MM cells, these drugs are ineffective in treating the cancer. As a result, MM death rates have remained relatively the same over the past 30 years with no significant breakthrough in treatment regimens (NIH; Figure 1).

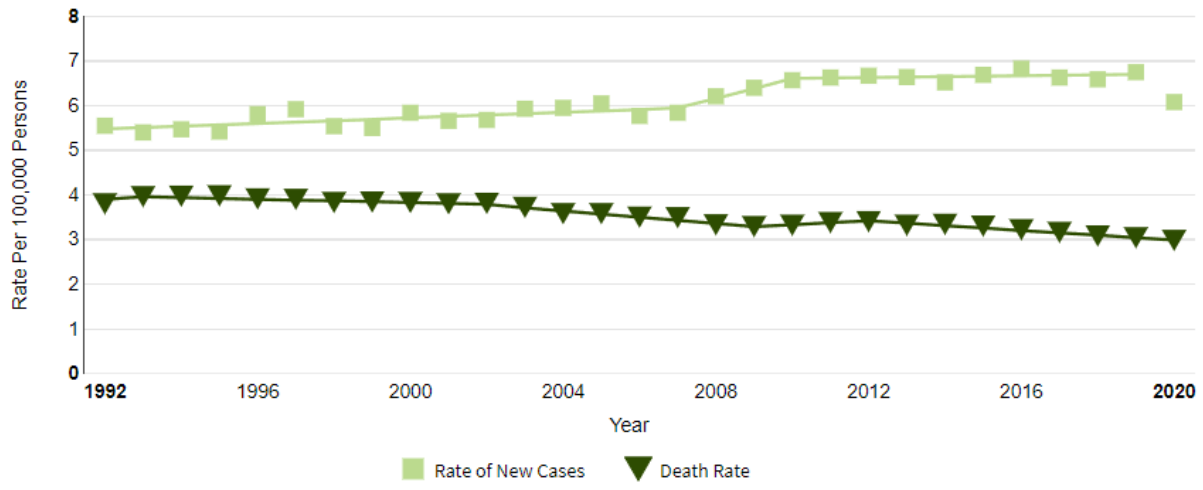


Figure 1: Observed rate of new cases (light green) and death rates (dark green) per 1000 people of multiple myeloma each year from 1992-2020 (NIH). Observed rate of new cases continue to steadily increase each year while death rates remain relatively constant. The increase in new cases since 1992 is 1.3% while death rates decreased by only 0.8%.

MM death rates have only decreased from 3.8% in 1992 to 3.0% in 2020, while new cases continue to rise. Even if this number seems small, the 5-year survival rate of MM (59.8%) is still lower than other common cancers such as breast cancer (90.8%) (NIH). Additionally,

proteasome inhibitors can cause dizziness, labored breathing, nausea, and numerous other side effects (Sharma and Preuss, 2022). As a result, there is a need to develop new, targeted therapies towards MM to improve survival rates and reduce adverse side effects.

1.3 Single Cell RNA Sequencing

Single-cell RNA sequencing (scRNA-seq) is a modern biotechnology which can be used to identify differentially expressed genes when comparing two different cells. The technique measures gene expression levels of thousands of genes in a single cell as whole integer values with each count representing every time the gene is expressed. Observing the differences in genetic expression for certain genes can provide valuable insight into which genes are overexpressed or underexpressed in cancerous MM cells. Using this information, scientists can then create targeted therapies which aim to control the expression of these genes.

scRNA-seq has shown potential to identify pivotal genes which cause cancer, as shown in Sultana et al. (2023). This paper utilized scRNA-seq data to identify 12 biomarkers and genes for non-small cell lung cancer such as MS4A1, CCL5, and GZMB. Furthermore, Ren et al. (2021) poses as yet another example of the capability to identify metastasis genes using scRNA-seq. The identification of the S100A4 gene expressed in tumor genes was found through scRNA-seq and proved to play a role in metastasis in future research. Therefore, utilizing scRNA-seq shows promising potential to identify specific genes that cause MM, which can ultimately provide a direction for future targeted therapy.

1.4 Identifying Driver Genes to Create Targeted Treatments

Targeting specific driver genes has made it possible to create more effective treatments for cancer. Chu et al. (2021) proposed the DKK1 gene to play a significant role in MM proliferation. The DKK1 gene is responsible for producing the DKK1 protein, which is an

inhibitor of the Wnt-β-Catenin pathway. The Wnt-β-Catenin pathway produces β-catenin, which is a key tumor cell proliferation regulator. Jiang et al. (2022) researchers further into this gene for its application in clinical therapy for MM. As a result of the discovery of DKK1 and its relation to MM, new drugs called DKK1 inhibitors have been created.

1.5 Machine Learning vs Current Approach

The most common approach to identify genetic differences between cell types is using the GEO2R program, which is a web tool that compares across groups and samples in the Gene Expression Omnibus (GEO) dataset. GEO2R uses exclusively GEO datasets and limma R packages to visualize, process, and perform statistical analyses. However, GEO2R does display numerous drawbacks such as being slow to analyze large datasets with many samples or genes, which is commonly characteristic of scRNA-seq datasets. There is a 10-minute cutoff for data processing which can prevent scRNA-seq analysis. In addition, GEO2R shows difficulty in analyzing raw or non-normalized data which limits the scope of what datasets can be used (U.S. National Library of Medicine).

On the other hand, machine learning shows promising potential to surpass current approaches in analyzing differentially expressed genes. Unlike GEO2R, machine learning models can process and excel in their predictions on much larger datasets. Additionally, machine learning models provide much more information in the features (Le et al., 2022). Machine learning is capable of detecting patterns in scRNA-seq datasets, which are complex due to thousands of genes being variables. This allows machine learning to potentially identify novel genes. Utilizing machine learning models to analyze scRNA-seq data could be useful in identifying novel genes that proliferate MM which have not been previously established in current analysis techniques.

1.6 Objectives

This project's goals are threefold: to create machine learning models to correctly identify the stage of MM progression a cell is in; identify signature genes for each MM stage; and to potentially identify novel genes which have not been identified in previous literature.

2. Methodology

2.1 Dataset

scRNA-seq data was obtained from Gene Expression Omnibus created by Boiarsky et al. 2022. The dataset included RNA expression values from 29,387 plasma cells (9,329 NBM, 817 MGUS, 8,431 SMM, and 10,790 MM) taken from 26 patients in varying stages of MM progression. Expression values of 22,088 genes from 4 different cell types (NBM, MGUS, SMM, MM) were measured.

2.2 Pre-Processing

All genes that had no expression values were removed from the dataset. Then, the XIST sex gene was removed as differential expression of this gene only results from gender, not because of the presence of MM (Boiarsky et al., 2022). Next, highly expressed genes that are not responsible for MM pathogenesis were removed. This included genes from the IGH, IGL, and IGK loci, as they are already known to be highly expressed antibodies in abnormal cells but have no importance in disease progression (Boiarsky et al., 2022). The dataset then was normalized by taking the log 2 of every expression value to reduce skew in the data for genes that are naturally expressed much more than others (Ilozumba et al., 2022). Log 2 normalization was chosen for this dataset because it is commonly used in differentially expressed gene datasets (Bergemann and Wilson, 2011). Highly variable genes were then selected using Scanpy because this study focuses on only observing genes that are differentially expressed in cell types. Highly variable

gene selection reduces the dimensionality of the dataset by removing genes with limited value (Boiarsky et al., 2022). Finally, the gene expression values were scaled to prepare for principal component analysis. Scaling standardizes the data by subtracting the mean from each respective value and scaling that value to its unit (0.0-1.0) variance (Abd El-Haleem et al., 2022).

2.3 Principal Component Analysis and Scree Plot

Principal component analysis (PCA) was utilized in this study to reduce high dimensionality scRNA-seq data (Figure 2). Because scRNA-seq data contains thousands of genes, it will be hard for machine learning models to account for all of these variables when making their classification decision. Therefore, PCA will condense these genes into principal component clusters, while still preserving as much information as possible. First, all genes are inputted into a covariance matrix to measure the relationship between every pair of genes. Next, eigenvectors and eigenvalues are computed to create principal components. Each principal component consists of genes with a different weight in how important it is. Principal components will then organize every cell to a certain cluster based on their gene expression values and find the notable genes which are responsible for creating different clusters (Jolliffe and Cadima 2016). Principal components were then inputted into machine learning models to determine which principal component was the most important in differentiating between cell types.

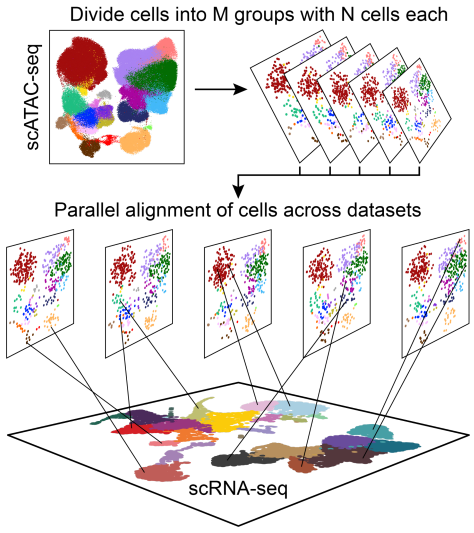


Figure 2: Schematic diagram of PCA utilized in single cell RNA Sequencing. All genes in the scRNA dataset is split into M principal components each with N number of genes. Cells can be mapped out into different clusters based on how similar their principal component are to each other.
Source: ArchR

Too little principal components may remove too much information while having too many can reduce model performance. To combat this issue, a scree plot was observed to determine the amount of principal components to keep (Figure 3). The scree plot maps out the proportion of variance explained by each principal component. At the point where explained variance levels off is the sufficient amount of principal components to keep. It was found that 30 principal components were sufficient for data analysis. Principal components were then inputted in machine learning models to differentiate between NBM, MGUS, SMM, and MM cell types (Michie et al., 2021). Principal components will then be listed with each gene ordered by most correlated to the cancer to least correlated.

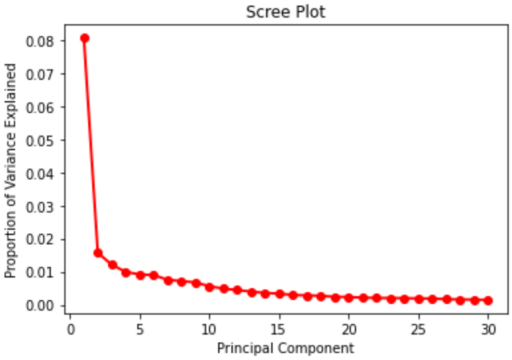


Figure 3: Scree plot which maps out each principal component and the proportion of variance it can explain. Explained variance levels out at around 30 principal components.

2.4 Models

This study utilized 3 different machine learning models: Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Classification accuracy for each model will be observed, and the most important principal component will be noted from each model.

Random Forest (RF) is a machine learning model that employs ensemble learning methods with a combination of decision trees fitted on randomly selected subsets of data (Breiman, 2001). To improve predictive ability and control for overfitting, RF averages the predictions from each decision tree to form its final prediction. This study utilized RF since ensemble learning methods have proven to be very successful in classification problems using gene expression data (Mahendran et al., 2020).

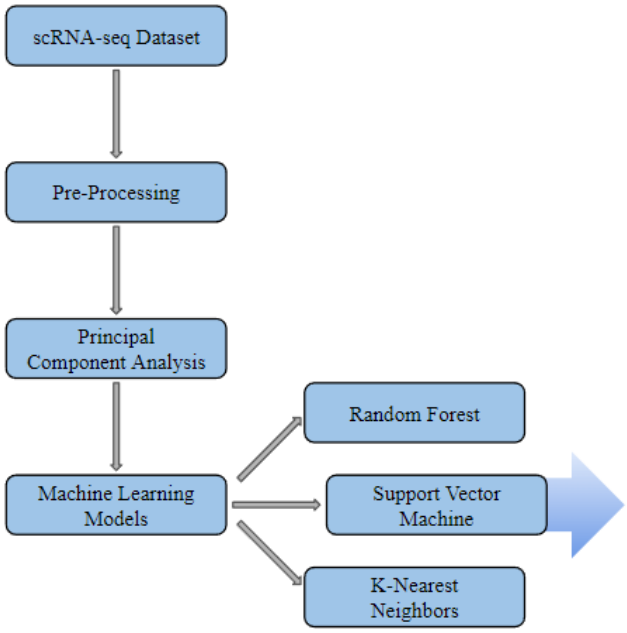
Support Vector Machine (SVM) is a machine learning model that makes its predictions by establishing a hyperplane (or decision boundary) that separates the data points from each class. This hyperplane is developed to be the farthest away from the support vectors—data points that are closest to other classes' data points—as possible. By projecting data points into a higher dimensional space, a process called the kernel trick, SVM is able to effectively improve its predictions (Huang et al., 2018). SVM was selected as a model as it has shown previous success (98% accuracy) in the classification of colon cancer with a gene expression dataset (Guyon et al., 2002).

K-Nearest Neighbors is a nonparametric learning algorithm in which the classification of an object depends on the values of its neighbors. Each datapoint assumes that its neighbors are indicators for its values and therefore weighs them more than distant values in its predictions (dos Santos Freitas et al., 2022). KNN has been previously shown to be highly accurate when tested on gene selection datasets (Mahendran et al., 2020).

2.5 Metrics

Classification metrics will determine how effective each model was in correctly identifying the current stage of MM progression in a cell. Metrics represent the model prediction out of all cell types in the dataset. For example, accuracy measures the amount of correct predictions out of all 4 cell types, then divides by the total number of cells. Categorical classification of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), true positive rate (TPR), and true negative rate (TNR) were observed. Accuracy, balanced accuracy, precision, recall, and F-1 score were the metrics included in this study to observe model performance. 80% of each cell type in the dataset was used for training each of the models while the other 20% was used for testing. Figure 4 displays the metrics in depth, along with the

entire workflow of the methodology. Macro-averaged scores calculate the scores from each class and take the unweighted mean of all classes. On the other hand, micro-averaged scores calculate the scores globally across classes and adjust for unbalanced data.



Metrics

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy: measures the number of correct predictions over the total number of predictions made

$$balanced\ accuracy = \frac{TPR + TNR}{2}$$

Balanced accuracy: the arithmetic mean between recall and specificity. It adjusts the accuracy score to be more representative of imbalanced datasets.

$$precision = \frac{TP}{TP+FP}$$

Precision: measures the number of correct positive predictions over the total number of positive predictions made.

$$recall = \frac{TP}{TP + FN}$$

Recall: measures the number of correct positive predictions over the total number of positive values.

$$F_1\ score = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

F-1 score: harmonic mean of precision and recall, indicating how well-balanced the predictions of the model are.

Figure 4: Overall flowchart of methodology. Describes equations used to calculate accuracy, balanced accuracy, precision, recall, and F-1 score.

2.6 Multicollinearity

Multicollinearity is when two or more independent variables show correlation with each other, making it difficult to distinguish their independent influence (Voss, 2005). Multicollinearity between identified driver genes would indicate that a change in expression levels of one gene was the result of a change in expression levels of another gene, making it unlikely that the gene expression levels changed due to MM. To ensure that the identified driver

genes were independent of each other, a multicollinearity test was run and the variance inflation factor (VIF) of each gene was calculated. A VIF value of 1 indicates no correlation between variables, and a VIF value of 10 or greater generally indicates multicollinearity (Kim, 2019). Identified driver genes that showed multicollinearity were eliminated from this study's analysis, as a change in gene expression is likely not the result of MM.

3. Results

3.1 Metrics Results

Table 1 lists the metric scores for RF, SVM, and KNN. RF was the overall best performing model, with the highest accuracy of 95.61%. RF outperformed the other two models in accuracy, precision, micro-averaged recall, and micro-averaged F-1 score (See Table 1). Scores are represented in decimal values and measure performance in each model in correctly identifying cell type. It should be noted that all three models performed extremely similarly to each other with accuracies differing by only 0.43%. Therefore, principal component importance will observe all three models, instead of just the best performing one. All three models will be observed to see which majority principal component was the most important.

Table 1: Table comparing metric scores between RF, SVM, and KNN. Metrics include accuracy, balanced accuracy, precision, recall, and F-1 score.

Model	Accuracy	Balanced Accuracy	Precision		Recall		F-1 Score	
			Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged	Macro-averaged	Micro-averaged
Random Forest	0.9561	0.8265	0.9473	0.9561	0.8265	0.9561	0.8610	0.9561
Support Vector Machine	0.9518	0.8431	0.9392	0.9518	0.8431	0.9518	0.8757	0.9518
K-Nearest Neighbors	0.9556	0.8510	0.9359	0.9556	0.8510	0.9556	0.8811	0.9556

3.2 Confusion Matrix

Model predictions were summarized with confusion matrices as shown in Figure 5. Highlighted blue boxes represent the model correctly identifying the right type of cell. All three models performed the worst in identifying MGUS cells, with many cells misidentified as NBM cells. However, it should be noted that there were significantly less MGUS cells in the dataset compared to every other cell type.

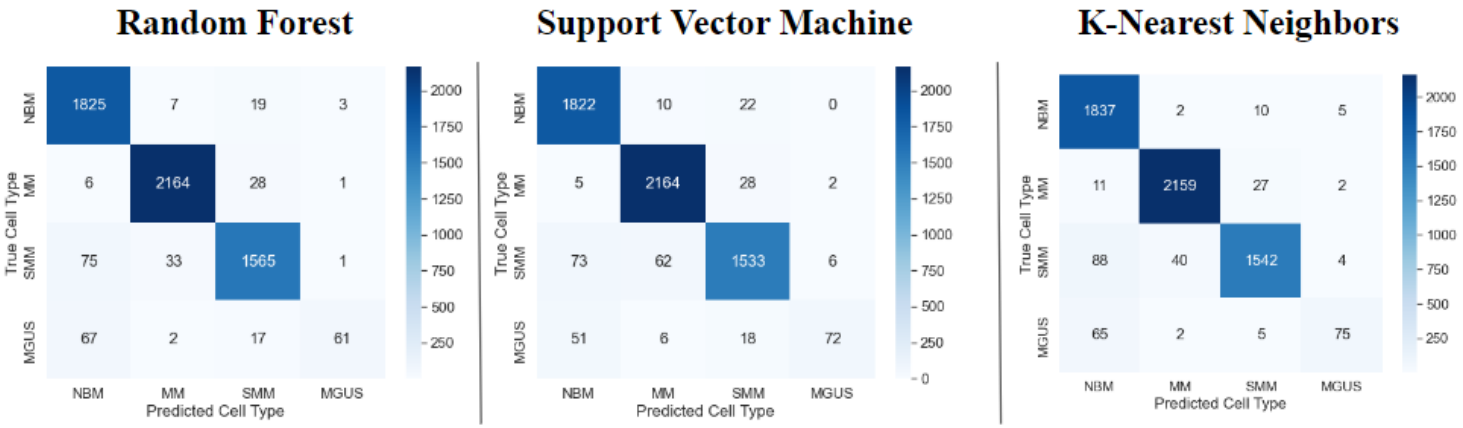


Figure 5: Confusion matrices representing prediction values compared between RF, SVM, and KNN. Predicted cell type decision by the model (X-axis) is put against true cell type (Y-axis). Each number represents a count for how many samples were predicted to by the model to be a specific cell type.

3.3 Feature Importance

Models then were evaluated to see which principal component was the most important in its classification decision. Each model used their respective algorithm to determine principal component importance and is illustrated in Figure 6. Importance score is plotted against principal components (PC) to compare between principal components. Note that 0 represents Principal Component 1 (PC1).

RF and SVM both determined that PC3 was the most important component in its classification decision, while KNN determined it to be PC5. Therefore, genes in the two most important components (PC3 and PC5) will be further discussed.

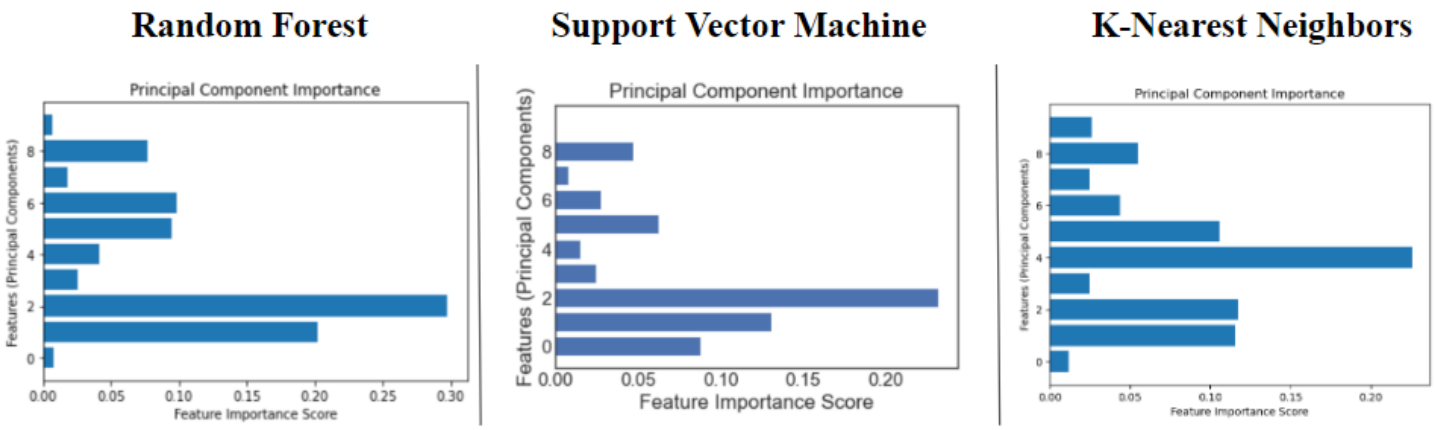


Figure 6: Principal Component Importance mapped against PC1-PC10 shown in RF, SVM, and KNN.

3.4 Genes

Listed in Table 2 are the top 5 most important genes in the first 9 principal components. These genes were weighed the most in each principal component, with the first gene being the most important. Only the first 9 principal components were used because the rest showed limited importance. Bolded genes will be further discussed later.

Table 2: Top 5 genes listed in top 9 principal components. PC3 was established to be the most important component by RF and SVM while PC5 was the most important in KNN.

PC1	PC2	PC3 (RF & SVM)	PC4	PC5 (KNN)
GAPDH	CST3	**ANKRD28**	LAG3	**HLA-DPA1**
RPL4	FRZB	NFKBIA	HBA2	HLA-DRB5
HNRNPA1	FCRLA	TMSB4X	HBA1	HLA-DRA
EEF2	**CH17-224D4.2**	H3F3B	PTP4A3	HLA-DQA1
ITM2C	WHSC1	CXCR4	LILRB4	HLA-DQA2

PC6	PC7	PC8	PC9
RP5-1171I10.5	STMN1	AC233755.2	AZGP1
SNHG25	NUSAP1	STMN1	AC233755.2
CH17-224D4.2	TOP2A	NTRK2	MAFB
CCL3	KIAA0101	NUSAP1	NTRK2
FGFR3	MKI67	TOP2A	P2RX1

3.5 Multicollinearity

The results of the multicollinearity test are listed in Table 3. Genes with VIF scores above 10 were highlighted and will be excluded because their expression levels are due to other genes, not MM itself. Only one gene, RPL4 (PC1) displayed a VIF above 10.

Table 3: Multicollinearity test using Variance Inflation Factor (VIF) scores. Genes with VIF scores above 10 were highlighted and excluded. Only one gene, RPL4 (PC1) was excluded.

PC1		PC2		PC3		PC4		PC5	
Genes	VIF	Genes	VIF	Genes	VIF	Genes	VIF	Genes	VIF
GAPDH	3.87	CST3	3.10	ANKRD28	2.26	LAG3	3.47	HLA-DPA1	3.66
RPL4	**15.03**	FRZB	2.84	NFKBIA	1.99	HBA2	5.17	HLA-DRB5	4.31
HNRNPA1	6.83	FCRLA	1.77	TMSB4X	1.59	HBA1	2.92	HLA-DRA	8.23
EEF2	7.26	CH17-224D4.2	2.28	H3F3B	2.21	PTP4A3	2.32	HLA-DQA1	5.26
ITM2C	2.80	WHSC1	1.92	CXCR4	1.93	LILRB4	1.39	HLA-DQA2	4.70
GLTSCR2	6.46	NPM1	2.81	LMNA	1.70	NUDT12	1.947	HLA-DPB1	2.764

PC6		PC7		PC8		PC9	
Genes	VIF	Genes	VIF	Genes	VIF	Genes	VIF
RP5-1171I10.5	1.68	STMN1	2.74	AC233755.2	3.13	AZGP1	1.67
SNHG25	1.86	NUSAP1	2.46	STMN1	1.91	AC233755.2	2.58
CH17-224D4.2	1.23	TOP2A	2.61	NTRK2	1.87	MAFB	2.08
CCL3	1.17	KIAA0101	2.18	NUSAP1	2.40	NTRK2	1.86
FGFR3	1.47	MKI67	2.13	TOP2A	2.80	P2RX1	1.25
CD1D	1.73	TYMS	2.33	MAFB	1.97	CYP20A1	1.42

4. Discussion

4.1 Design Choices

The overall design of this study aimed to develop a machine learning model to classify MM cells and then to determine which genes were responsible for the classification decision. MM was chosen because it is a prevalent cancer worldwide, with still a relatively low 5-year survival rate compared to other cancers. Additionally, MM still lacks a specific treatment regimen that targets genetic pathways characteristic of the cancer. Therefore, this study aimed to identify genes that progress the cancer so scientists can further research these genes for targeted treatment.

All of the models achieved above a 95% accuracy in classifying each of the 4 stages of MM progression. Every model performed the best in identifying MM cells and the worst at identifying MGUS cells. This study hypothesizes that the models were the worst in identifying MGUS cells because their genetic composition was similar to NBM cells. This explains why many MGUS cells were misidentified to be NBM. As shown in previous literature by Rajkumar (2022), MGUS is still a benign stage in MM progression and M proteins are not present yet. MGUS is the most similar stage to NBM so genetic differences may not be apparent.

4.2 Genes

All genes listed in the table are hypothesized to play a role in MM progression. Genes were cross-validated in an extensive PubMed database search to determine if PCA and the models worked correctly. If the gene was previously established in literature to be associated with MM progression, it can help validate the possible genes that are implicated in progressing MM, which can then be extended to novel genes found in this study.

ANKRD28, the most important gene in PC3, which is also the most important PC determined by both RF and SVM, has previously been established to be responsible for genetic disorders when mutated (Kiyokawa E et al., 2009). ANKRD28 plays a role in cellular adhesion and promotes migration which are two factors important in cancer proliferation. Furthermore, ANKRD28 is found to be hypomethylated in MGUS cells as shown by Heuck et al. (2013). Hypomethylation of the gene can be a possible explanation for why gene expression levels were higher in cancerous cells compared to NBM cells. Although the exact function of ANKRD28 is unclear, this gene still shows potential to be a promising indicator for progression and identification of MGUS cells (Wu et al., 2023).

HLA-DPA1, the most important gene in PC5, which is the most important PC determined by KNN, is responsible for hypoxia in MM when mutated (Yang et al., 2020). Yang et al. states that this gene can be a potential indicator with prognostic values in multiple myeloma, which needs to be further investigated.

Most notably, two genes listed in the table, RP5-1171I10.5 and CH17-224D4.2, were not yet established to be associated with MM. RP5-1171I10.5 is responsible for the production of a long non-coding RNA which is responsible for cell proliferation in breast cancer. This gene's role in breast cancer progression shows potential to also be important in other types of cancer such as MM. Unfortunately, the function of the CH17-224D4.2 gene has not yet been reported in PubMed so it will be excluded from the study.

4.3 Potential for application

All three models achieved a promising accuracy above 95%, which shows its potential to be effective in diagnosing MM in plasma cells. These models can be implemented in clinical studies to swiftly identify MM when given genetic expression data.

More importantly, this study discovered the novel gene, RP5-1171I10.5, to be associated with MM progression. Based on the models' performance and cross-validation, this gene is likely to play a role in MM. This gene should be further studied to create more specific targeted genetic therapies for MM.

5. Limitations and Future Research

5.1 Limitations

One limitation of this study is that the genes identified by the machine learning models may be associated with MM, but the data does not allow for a conclusion that change in the expression of those genes causes MM. Therefore, this study was only able to identify potential driver genes of MM that require further research.

Imbalances in the scRNA-seq dataset was another limitation in this study. MGUS had a significantly lower number of samples compared to the other stages of MM progression, leading to the machine learning models showing difficulty in classifying MGUS cells. Additionally, the patients sampled in the dataset were mostly white; however, MM has shown to be around two to three times as common in African American patients compared to non-Hispanic white patients (Dong et al., 2022). Differentiations in genetics between races may have skewed the data and resulted in identified genes that are not representative of the general population.

Lastly, pre-existing medical conditions in sampled patients could have led to changes in the expression of certain genes that are not associated with MM, but instead that pre-existing condition. Therefore, the gene expression values in the data could have been skewed, leading to falsely identified genes.

5.2 Future Applications and Research

This study only utilized three different machine learning models and the results showing the trend of the most important principal components was unclear, as RF and SVM identified PC3 as the most important while KNN identified PC5 as the most important. In the future, this study could implement additional machine learning models such as CNN, LSTM, and GNN—this would allow for the identification of a clearer trend in the most important principal components. Additionally, the methodology used in this study could be applied to different types of cancer or to different conditions that have scRNA-seq data.

A long-term goal of this research would be to conduct a sub-analysis of the patients sampled in the dataset—patient characteristic and diagnostic factors were provided by Boiarsky et al. (2022) in a supplementary data file. Assessing patient characteristics would allow this study to analyze how factors such as age, race, sex, and ethnicity would impact the genes identified to be associated with MM. Diagnostic factors such as if patients were treated during MGUS or SMM, M protein levels, and days until a follow-up could be analyzed for associated with changes in the expression levels of certain genes. Additionally, identified genes could be further studied using model organisms such as mice. Gene biotechnology will be used to knockout or manipulate each identified gene to observe its significance in cellular homeostasis. Eventually, new therapeutics can be made which modify each specific identified gene back to its normal expression levels.

6. Conclusion

The main objective of this study was to classify whether a cell is in a stage of MM, and to then identify driver genes for MM. All models achieved above a 95% accuracy in identifying the correct type of cell. Genes listed in the principal components were consistent with previous

literature, which indicates that the models show promising potential to correctly identify driver genes. RP5-1171I10.5 was a novel gene discovered in this study which has not yet been established to be associated with MM. Based on its function in breast cancer, the function of RP5-1171I10.5 needs to be further studied with its relation in MM. Ultimately, targeting driver genes in therapies can improve survival rates and create a much better prognosis of MM.

7. Acknowledgments

This study would like to express gratitude to Ms. Spinelli, who made this project possible and advised us along the steps to create this project.

8. Bibliography

Abd el-haleem, A. M., Mohamed, N. E.-D. M., Abdelhakam, M. M., & Elmesalawy, M. M. (2022). A machine learning approach for movement monitoring in clustered workplaces to control covid-19 based on geofencing and fusion of wi-fi and magnetic field metrics. *Sensors*, *22*(15), 5643. https://doi.org/10.3390/s22155643

Albagoush SA, Shumway C, Azevedo AM. Multiple Myeloma. 2023 Jan 30. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan–. PMID: 30521185.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://link.springer.com/article/10.1023/a:1010933404324

Bergemann, T. L., & Wilson, J. (2011). Proportion statistics to detect differentially expressed genes: A comparison with log-ratio statistics. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-228

Boiarsky, R., Haradhvala, N. J., Alberge, J.-B., Sklavenitis-pistofidis, R., Mouhieddine, T. H., Zavidij, O., Shih, M.-C., Firer, D., Miller, M., El-khoury, H., Anand, S. K., Aguet, F., Sontag, D., Ghobrial, I. M., & Getz, G. (2022). Single cell characterization of myeloma

and its precursor conditions reveals transcriptional signatures of early tumorigenesis. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-33944-z

Cardoso-fernandes, J., Teodoro, A. C., Lima, A., & Roda-robles, E. (2020). Semi-Automatization of support vector machines to map lithium (Li) bearing pegmatites. *Remote Sensing*, *12*(14), 2319-32. https://doi.org/10.3390/rs12142319

Chu, H. Y., Chen, Z., Wang, L., Zhang, Z.-K., Tan, X., Liu, S., Zhang, B.-T., Lu, A., Yu, Y., & Zhang, G. (2021). Dickkopf-1: A promising target for cancer immunotherapy. Frontiers in Immunology, 12. https://doi.org/10.3389/fimmu.2021.658097

Das, P., Roychowdhury, A., Das, S., Roychoudhury, S., & Tripathy, S. (2020). SigFeature: Novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00247

Das, S., Juliana, N., Yazit, N. A. A., Azmani, S., & Abu, I. F. (2022). Multiple myeloma: Challenges encountered and future options for better treatment. *International Journal of Molecular Sciences*, *23*(3), 1649. https://doi.org/10.3390/ijms23031649

Dong, J., Garacci, Z., Buradagunta, C. S., D'Souza, A., Mohan, M., Cunningham, A., Janz, S., Dhakal, B., Thrift, A. P., & Hari, P. (2022). Black patients with multiple myeloma have better survival than white patients when treated equally: a matched cohort study. Blood cancer journal, 12(2), 34. https://doi.org/10.1038/s41408-022-00633-5

Dos santos freitas, M. M., Barbosa, J. R., Dos santos martins, E. M., Da silva martins, L. H., De souza farias, F., De fátima henriques lourenço, L., & Da silva e silva, N. (2022). KNN

algorithm and multivariate analysis to select and classify starch films. *Food Packaging and Shelf Life*, *34*, 100976. https://doi.org/10.1016/j.fpsl.2022.100976

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.m., Piñeros, M., Znaor, A., & Bray, F. (2018). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, *144*(8), 1941-1953. https://doi.org/10.1002/ijc.31937

Hosen, N., Ichihara, H., Mugitani, A., Aoyama, Y., Fukuda, Y., Kishida, S., Matsuoka, Y., Nakajima, H., Kawakami, M., Yamagami, T., Fuji, S., Tamaki, H., Nakao, T., Nishida, S., Tsuboi, A., Iida, S., Hino, M., Oka, Y., Oji, Y., & Sugiyama, H. (2011). CD48 as a novel molecular target for antibody therapy in multiple myeloma. *British Journal of Haematology*, *156*(2), 213-224. https://doi.org/10.1111/j.1365-2141.2011.08941.x

Heuck CJ, Mehta J, Bhagat T, Gundabolu K, Yu Y, Khan S, Chrysofakis G, Schinke C, Tariman J, Vickrey E, Pulliam N, Nischal S, Zhou L, Bhattacharyya S, Meagher R, Hu C, Maqbool S, Suzuki M, Parekh S, Reu F, Steidl U, Greally J, Verma A, Singhal SB. Myeloma is characterized by stage-specific alterations in DNA methylation that occur early during myelomagenesis. J Immunol. 2013 Mar 15;190(6):2966-75. doi: 10.4049/jimmunol.1202493. Epub 2013 Feb 13. PMID: 23408834; PMCID: PMC4581585.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer genomics & proteomics*, *15*(1), 41–51. https://doi.org/10.21873/cgp.20063

Hwang, B., Lee, J. H., & Bang, D. (2021). Author correction: Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *53*(5), 1005. https://doi.org/10.1038/s12276-021-00615-w

Ilozumba, M. N., Yao, S., Llanos, A. A. M., Omilian, A. R., Zhang, W., Datta, S., Hong, C.-C., Davis, W., Khoury, T., Bandera, E. V., Higgins, M., Ambrosone, C. B., & Cheng, T.-Y. D. (2022). MTOR pathway gene expression in association with race and clinicopathological characteristics in black and white breast cancer patients. *Discover Oncology*, *13*(1). https://doi.org/10.1007/s12672-022-00497-y

Jiang, H., Zhang, Z., Yu, Y., Chu, H. Y., Yu, S., Yao, S., Zhang, G., & Zhang, B.-T. (2022). Drug discovery of dkk1 inhibitors. Frontiers in Pharmacology, 13. https://doi.org/10.3389/fphar.2022.847387

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kanapuru, B., Fernandes, L. L., Fashoyin-aje, L. A., Baines, A. C., Bhatnagar, V., Ershler, R., Gwise, T., Kluetz, P., Pazdur, R., Pulte, E., Shen, Y.-L., & Gormley, N. (2022). Analysis of racial and ethnic disparities in multiple myeloma US FDA drug approval trials. *Blood Advances*, *6*(6), 1684-1691. https://doi.org/10.1182/bloodadvances.2021005482

Kawano, Y., Zavidij, O., Park, J., Moschetta, M., Kokubun, K., Mouhieddine, T. H., Manier, S., Mishima, Y., Murakami, N., Bustoros, M., Pistofidis, R. S., Reidy, M., Shen, Y. J., Rahmat, M., Lukyanchykov, P., Karreci, E. S., Tsukamoto, S., Shi, J., Takagi, S., . . . Ghobrial, I. M. (2018). Blocking ifnar1 inhibits multiple myeloma–driven treg expansion

and immunosuppression. *Journal of Clinical Investigation*, *128*(6), 2487-2499. https://doi.org/10.1172/JCI88169

Khozeimeh, F., Sharifrazi, D., Izadi, N. H., Joloudari, J. H., Shoeibi, A., Alizadehsani, R., Tartibi, M., Hussain, S., Sani, Z. A., Khodatars, M., Sadeghi, D., Khosravi, A., Nahavandi, S., Tan, R.-S., Acharya, U. R., & Islam, S. M. S. (2022). RF-CNN-F: Random forest with convolutional neural network features for coronary artery disease diagnosis based on cardiac magnetic resonance. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-15374-5

Kim, J., Xu, Z., & Marignani, P. A. (2021). Single-cell RNA sequencing for the identification of early-stage lung cancer biomarkers from circulating blood. *Npj Genomic Medicine*, *6*(1). https://doi.org/10.1038/s41525-021-00248-y

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558-569. https://doi.org/10.4097/kja.19087

Kiyokawa E, Matsuda M. Regulation of focal adhesion and cell migration by ANKRD28-DOCK180 interaction. Cell Adh Migr. 2009 Jul-Sep;3(3):281-4. doi: 10.4161/cam.3.3.8857. Epub 2009 Jul 27. PMID: 19458477; PMCID: PMC2712811.

Kyle RA, Rajkumar SV. Management of monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM). Oncology (Williston Park). 2011 Jun;25(7):578-86. PMID: 21888255; PMCID: PMC3923465.

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

Mateos, M.-I., Kumar, S., Dimopoulos, M. A., González-calle, V., Kastritis, E., Hajek, R., De larrea, C. F., Morgan, G. J., Merlini, G., Goldschmidt, H., Geraldes, C., Gozzetti, A., Kyriakou, C., Garderet, L., Hansson, M., Zamagni, E., Fantl, D., Leleu, X., Kim, B.-S., . . . San-miguel, J. (2020). International myeloma working group risk stratification model for smoldering multiple myeloma (SMM). *Blood Cancer Journal*, *10*(10). https://doi.org/10.1038/s41408-020-00366-3

Rajkumar, S. V. (2016). Multiple myeloma: 2016 update on diagnosis, risk-stratification, and management. *American Journal of Hematology*, *91*(7), 719-734. https://doi.org/10.1002/ajh.24402

Rajkumar, S. V. (2022). Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *American Journal of Hematology*, *97*(8), 1086-1107. https://doi.org/10.1002/ajh.26590

Rajkumar, S. V. (2022). Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *American Journal of Hematology*, *97*(8), 1086-1107. https://doi.org/10.1002/ajh.26590

Rukhsar, L., Bangyal, W. H., Ali khan, M. S., Ag ibrahim, A. A., Nisar, K., & Rawat, D. B. (2022). Analyzing rna-seq gene expression data using deep learning approaches for cancer classification. *Applied Sciences*, *12*(4), 1850. http://dx.doi.org/10.3390/app12041850

Sharma A, Preuss CV. Bortezomib. [Updated 2022 Sep 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519559/

Song, D., Yuan, X., Li, Q., Zhang, J., Sun, M., Fu, X., & Yang, L. (2023). Intrusion detection model using gene expression programming to optimize parameters of convolutional neural network for energy internet. *Applied Soft Computing*, *134*, 109960. https://doi.org/10.1016/j.asoc.2022.109960

Sultana, A., Alam, M. S., Liu, X., Sharma, R., Singla, R. K., Gundamaraju, R., & Shen, B. (2023). Single-cell rna-seq analysis to identify potential biomarkers for diagnosis, and prognosis of non-small cell lung cancer by using comprehensive bioinformatics approaches. *Translational Oncology*, *27*, 101571. https://doi.org/10.1016/j.tranon.2022.101571

Survarachakan, S., Prasad, P. J. R., Naseem, R., Pérez de frutos, J., Kumar, R. P., Langø, T., Alaya cheikh, F., Elle, O. J., & Lindseth, F. (2022). Deep learning for image-based liver analysis — A comprehensive review focusing on malignant lesions. *Artificial Intelligence in Medicine*, *130*, 102331. https://doi.org/10.1016/j.artmed.2022.102331

Van de donk, N. W. C. J., Pawlyn, C., & Yong, K. L. (2021). Multiple myeloma. *The Lancet*, *397*(10272), 410-427. https://doi.org/10.1016/S0140-6736(21)00135-5

Voss, D. S. (2005). Multicollinearity. https://doi.org/10.1016/B0-12-369398-5/00428-X

Wu, Y., Meng, L., Cai, K., Zhao, J., He, S., Shen, J., Wei, Q., Wang, Z., Sooranna, S., Li, H., & Song, J. (2021). A tumor-infiltration cd8+ T cell-based gene signature for facilitating the prognosis and estimation of immunization responses in hpv+ head and neck squamous cell cancer. Frontiers in Oncology, 11. https://doi.org/10.3389/fonc.2021.749398

Yang, J., Wang, F. & Chen, B. HLA-DPA1 gene is a potential predictor with prognostic values in multiple myeloma. BMC Cancer 20, 915 (2020). https://doi.org/10.1186/s12885-020-07393-0