Veedhi Solanki

Department of Bioinformatics, University of Guelph

Statistical Bioinformatics

March 18, 2024

PROBLEM 1

Predicting COVID-19 Disease Progression Using Elastic-Net Regularized Logistic Regression

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has emerged as a significant global health crisis. A notable characteristic of severe cases of COVID-19 is the cytokine storm, a hyperinflammatory condition marked by an excessive immune response. This response is characterized by elevated levels of pro-inflammatory cytokines in the blood, correlating with disease severity, complications, and, in some cases, fatal outcomes. Understanding the dynamics of cytokine levels and their association with disease progression is crucial for identifying potential therapeutic targets and improving patient management strategies. This report aims to develop a predictive model using the Elastic-Net regularization framework to analyze cytokine profiles and patient characteristics, thereby predicting COVID-19 disease progression. The Elastic-Net approach is chosen for its ability to handle the multicollinearity among cytokines and to select relevant predictors by combining the properties of both Ridge and Lasso regression techniques.

Data Preparation and Preprocessing

In the initial phase of our analysis, we begin with preparing and preprocessing the dataset to ensure its preparedness for developing a predictive model. This process involved several critical steps, each contributing to the integrity and usability of the data for further analysis. We began by loading essential R libraries: `glmnet` for regression analysis, `pROC` for ROC curve analysis, `caret` for general data preparation and model evaluation, `readxl` for reading Excel files, and `ggplot2` for data visualization. These libraries provide a comprehensive toolkit for data manipulation, analysis, and visualization, catering to our analytical needs.

The dataset, titled "Immunologic profiles of patients with COVID-19.xlsx," was imported into R. This dataset comprises measurements of various cytokines, chemokines, and acute phase proteins from plasma samples of 76 patients who tested positive for COVID-19. These measurements are pivotal for understanding the immune response in COVID-19 patients and predicting disease progression. Our preliminary analysis of the dataset revealed two categorical variables: SEX and Severity. The SEX variable, indicating the gender of the patients, was binary-encoded to facilitate analysis, with males represented as 1 and females as 0. Similarly, the Severity variable, reflecting the disease's severity, was encoded numerically to distinguish between mild and severe cases, with severe cases marked as 1.

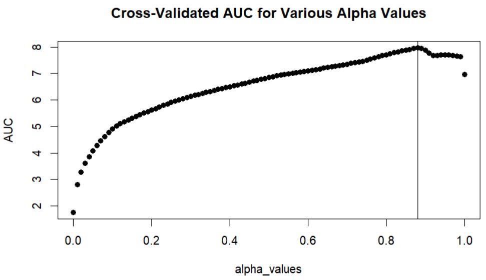
A notable observation during this phase was the data imbalance in both SEX (50 males to 26 females) and Severity (44 severe cases against 32 mild ones) variables. This imbalance could potentially influence the model's performance and was thus an important factor to consider in the analysis. We conducted a thorough examination for missing values across the dataset and found no missing values, eliminating the need for imputation strategies. This absence of missing data suggests a well-curated dataset, allowing for a more straightforward analysis. The 'Patient Number' column, seemed unnecessary for model development, was removed from the dataset. This decision was based on the assumption that patient identifiers do not contribute to predictive modeling of disease progression.

Furthermore, we split the dataset into a training set (75%) and a test set (25%) to enable model training and evaluation. This split was performed using a reproducible random seed (1717) to ensure consistency in analysis. The final step in our data preparation process involved normalizing the training and test datasets to center and scale the variables. This normalization is crucial for models like Elastic Net, which are sensitive to the scale of the input features. By normalizing the data, the aim was to improve the model's convergence and performance. The dataset was then converted to matrix form, as required by the `glmnet` function, marking the completion of our data preparation and preprocessing phase. With a clean, normalized dataset, we moved to the model development stage, aiming to predict COVID-19 disease progression based on cytokine profiles and patient characteristics.

Initial Model Selection

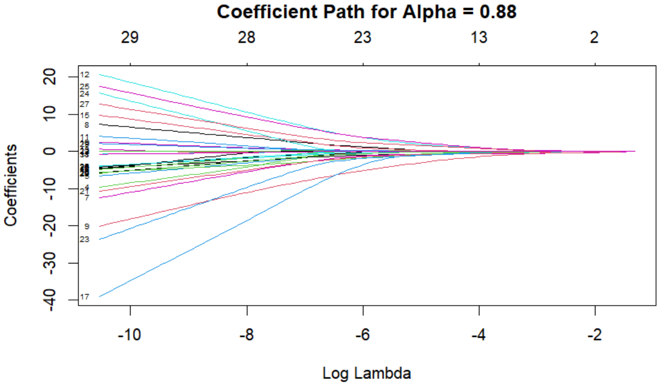
In the pursuit of constructing a predictive model for COVID-19 disease progression, we used Elastic-Net regularization, a method that balances the dimensionality reduction of Lasso (L1 regularization) with the feature-coefficient shrinkage characteristic of Ridge (L2 regularization). A range of alpha values, from 0 to 1 in increments of 0.01, was evaluated to ascertain the best model via 10-fold cross-validation. The fold IDs were meticulously generated and preserved for grid search consistency.

The plot on the right, the Cross-Validated AUC (Area Under the Curve) graph, shows how alpha values rise and eventually plateau. This particular trend suggests an increased enhancement in the model's predictive performance as we shift from Ridge-dominant regularization towards a Lasso-dominant approach. The graph culminates at an alpha of 0.88, beyond which the AUC stabilizes, signaling that further increase in the Lasso penalty does not translate to substantial gains in model performance.

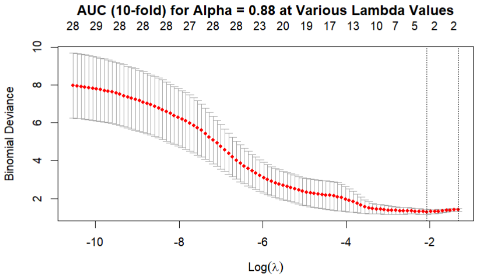


The alpha value of 0.88 emerged as the optimal parameter, as it corresponds to the peak of the AUC at approximately 7.97. This figure is significant as it underscores a preference for a model that incorporates more of the Lasso's feature selection capabilities without forgoing the Ridge's strength in addressing multicollinearity among predictors. Therefore, at this alpha, the Elastic-Net method optimizes for a minimal representation that maintains predictive reliability.

The identification of the 'best alpha' was further substantiated by evaluating the corresponding model's coefficient path as in the plot on the right, where the trajectory of each coefficient's magnitude was plotted against the log-transformed lambda values.



The model's complexity is fine-tuned with varying degrees of penalty strength (lambda), with the optimal lambda identified as 0.1266807 when alpha is fixed at 0.88. This lambda value represents the point of minimal deviance without compromising the model's predictive ability, as evidenced by the low binomial deviance in the AUC plot.

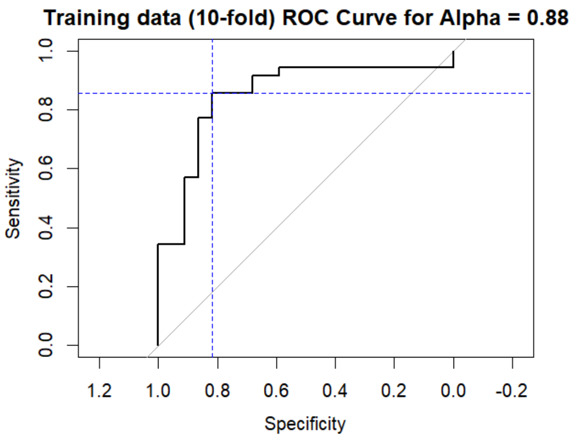


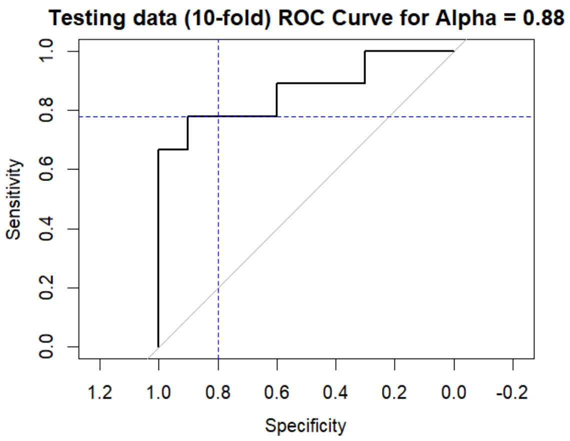
The predictive modeling process undertaken here begins with the careful consideration of training and testing data to evaluate model performance using Elastic Net regularization, a technique that encompasses both Lasso and Ridge regularizations, controlled by the alpha parameter. In training, the model selection relies on a comprehensive grid search over a continuum of alpha values, with the performance measured by the area under the ROC curve (AUC). The AUC offers an aggregate measure of model performance across all classification thresholds.

The model's proficiency in distinguishing between the two states of disease progression—non-severe and severe—is visually and numerically articulated through the ROC curve, and substantiated by the AUC values. During training, an AUC of approximately 0.85 was achieved at 10-fold cross validation, and this was closely matched by the AUC on the test data, which was around 0.87, indicating a model with good generalization capabilities.

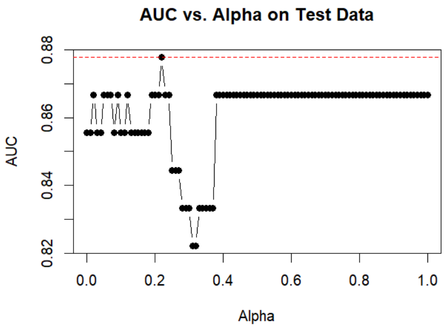
Additionally, sensitivity and specificity are optimized using the min-max approach, which seeks to maximize the minimum of the two metrics. This results in a balanced trade-off between the true positive rate (sensitivity) and the true negative rate (specificity), enhancing the model's diagnostic ability.

Additionally, the visualizations provided, such as the AUC vs. Alpha plot on the test data, facilitate a deeper understanding of the model's performance across various degrees of regularization. This aids in understanding a model's predictive performance on real data. Notably, for the test set, the model with an alpha of 0.22 has 0.1 difference in AUC with an alpha of 0.88, suggesting that a slight deviation towards Ridge regularization (since Lasso corresponds to alpha close to 1) could be beneficial for this specific dataset.





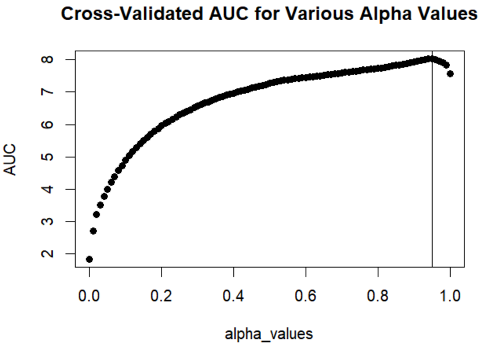
The plots above with ROC curves for training and testing, with their respective AUC values, show a satisfactory level of sensitivity and specificity, indicating the model's adeptness at classification tasks. Meanwhile, the AUC vs. Alpha plot encapsulates the model's stability across different alpha values, with the optimal alpha identified where the AUC peaks.



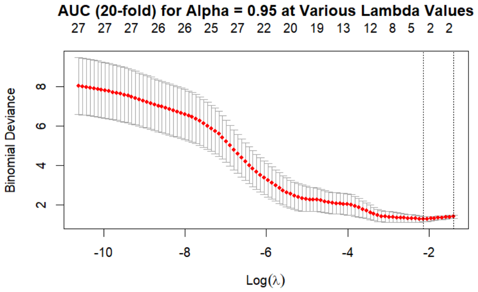
In summary, the analytical strictness applied here, underpinned by statistical validation techniques and graphically represented results, confirms the robustness of the Elastic Net model chosen. With AUC values that are impressive for both the training and test datasets, and an optimal threshold that ensures a balanced sensitivity-specificity trade-off, the model stands as a potent tool for predicting COVID-19 disease progression, while also allowing for interpretability and insight into the influence of various cytokines on disease severity.

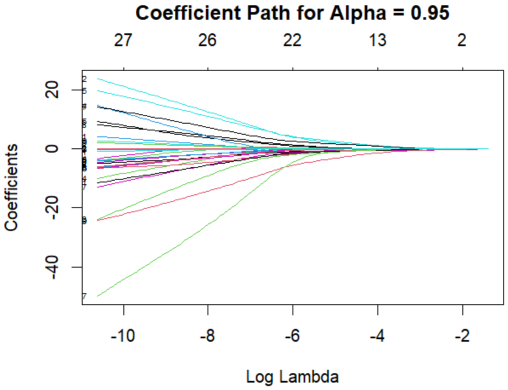
20-fold Cross-Validation

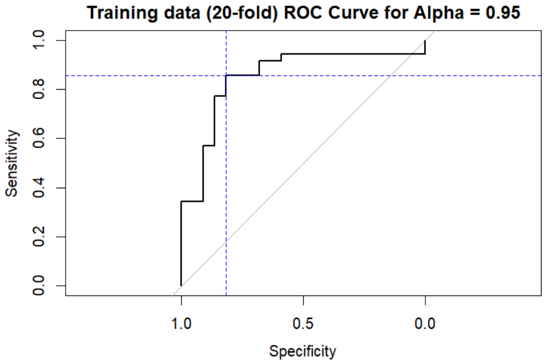
Shifting our focus to a 20-fold cross-validation, we examined the stability and reliability of the model selection process. The best alpha identified was 0.95 as shown in the plot on the right, which is a slight shift towards a more Lasso-predominant model. This shift is observed in the cross-validated AUC plot, where the AUC peaked slightly higher than in the 10-fold scenario at AUC 8.03.

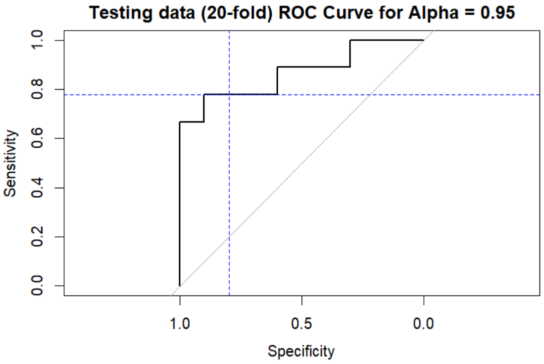


The coefficient path and ROC curve analysis and plots for 20-fold shown below for training and testing data closely mirrored those from the 10-fold validation, maintaining a high degree of model consistency. The training and testing ROC curves and sensitivity/specificity metrics were consistent with the 10-fold results, with AUCs of 0.849 and 0.867, respectively, and the decision threshold for the test data was slightly higher at 0.6421231.









A direct comparison shows that the 20-fold cross-validation resulted in a higher peak AUC value and a marginally different alpha value of 0.95, compared to an alpha of 0.88 from the 10-fold. Despite the increased computational demand of the 20-fold cross-validation, the slight improvement in AUC (from approximately 7.97 to 8.03) did not result in substantial changes in the model's predictive accuracy, as indicated by similar AUC values on the test data (0.867 for both).

The comparative analysis demonstrates that while 20-fold cross-validation may provide a slightly more accurate estimate of model performance due to a more comprehensive error estimation, it may not always translate into a significantly better model when weighed against the additional computational resources required. The model's stability across both validation methods, indicated by the consistent AUC values, suggests that either cross-validation scheme can be used effectively for this data set.

Cytokine Influence on COVID-19 Progression and Severity

The elastic net model identified PTX3 and IL-6 as significant cytokines with notable negative coefficients, indicating their importance in the prediction of COVID-19 disease progression among patients (Giamarellos-Bourboulis et al., 2020; Lucas et al., 2020). PTX3, with a coefficient of -0.3520082, and IL-6, with a coefficient of -0.2645898, emerged as key predictors. Their association with severe disease outcomes corroborates the hyperinflammatory state characterized by a cytokine storm, a critical aspect of COVID-19 pathology (Blanco-Melo et al., 2020; McElvaney et al., 2018).

Association of Cytokines with COVID-19 Disease Severity and significant cytokines

Upon closer examination of individual cytokines, IL-6 and PTX3 also displayed the most significant p-values (0.0002124993 and 0.0002901263, respectively), given their significant negative coefficients in the model and their low p-values, suggesting a high degree of confidence in their predictive power (Coomes & Haghbayan, 2020; Del Valle et al., 2020). It also reaffirms their strong association with COVID-19 disease severity. IL-6's role in activating autoimmunity and chronic inflammation is well-documented, and its elevated levels are often linked to severe disease. Similarly, PTX3 is involved in inflammatory responses and is indicative of tissue inflammation, which is prevalent in severe COVID-19 cases.

In addition to PTX3 and IL-6, the elastic net model identified other cytokines with statistical significance in predicting COVID-19 disease progression (Huang et al., 2020):

- CD27 (Coefficient: -0.0007754, P-Value: 0.0129612853) is implicated in the activation of B-cells. Its significant role in the immune response to COVID-19 suggests a potential impact on the progression of the disease, possibly through B-cell-mediated immunity.
- TNF-α (Coefficient: -0.02525736, P-Value: 0.0187304178) is a key proinflammatory cytokine. Its elevated levels are known to contribute to systemic inflammation, a characteristic of severe COVID-19 cases.
- HVEM (Coefficient: -0.0001999, P-Value: 0.0426783801) and CD28 (Coefficient: -0.02383565, P-Value: 0.0461238636) serve as co-signaling molecules in the immune system, influencing the activation and proliferation of T cells, which are critical in the body's response to viral infections like SARS-CoV-2.
- PD-1 and CTLA-4, with coefficients of -0.01644374 and -0.03522158, respectively, are inhibitory receptors that regulate the immune system, suggesting that modulation of immune checkpoints may be associated with disease severity.

Correlation of Patient Age with COVID-19 Disease Severity

In addition to cytokines, the patient's age was found to have a significant negative association with disease severity (p-value: 0.0159115167), contradicting the general observation where advanced age is a risk factor for severe disease. However, this may be due to the imbalance in the age data as there is comparatively more data for ages over 60 than for younger age categories. For instance, there is only a single observation for patients under 40 which can skew results and affect the trend. Overall, this could point to a particular demographic or clinical profile of the patient cohort in the dataset that warrants further investigation.

Model Performance: 10-Fold vs. 20-Fold Cross-Validation

The evaluation of the model's performance in classifying COVID-19 disease severity was undertaken through both 10-fold and 20-fold cross-validation approaches, with particular attention to accuracy, sensitivity, specificity, and precision metrics. Remarkably, the performance metrics from both the 10-fold and 20-fold cross-validation did not display substantial differences.

For the 10-fold cross-validation, we observed:

- Accuracy: 0.842105
- Sensitivity: 0.875000
- Specificity: 0.818182
- Precision: 0.777778

The 20-fold cross-validation yielded the following:

- Accuracy: 0.842105
- Sensitivity: 0.875000
- Specificity: 0.818182
- Precision: 0.777778

The mirrored results in performance metrics indicate that increasing the fold count from 10 to 20 did not alter the model's predictive capability, as evidenced by the consistent AUC values (approximately 0.85 for training and 0.87 for testing across both folds). This consistency suggests that the model is stable and the data is homogeneously informative, which allows for reliable cross-validation regardless of the fold count.

Implications and Conclusions

The analysis signifies that increasing the granularity of cross-validation, from 10-fold to 20-fold, did not lead to appreciable changes in the predictive performance of the model. This implies that the model selection is robust and the data possesses low variability in the context of the predictive task at hand. Additionally, this serves as an affirmation of the reproducibility and reliability of the chosen model.

The examination of cytokines and their association with COVID-19 progression and severity underscores the importance of a multifactorial approach in understanding the disease. The identification of significant cytokines like IL-6 and PTX3 contributes to a deeper comprehension of the immune response in COVID-19, offering potential biomarkers for disease progression and severity.

Problem 2

Perform PCA on this dataset treating differentially expressed genes as the variables. Visualize the PCs and investigate how they relate to cell and status types. Write down your interpretation considering the conclusions drawn in the article (Holmes et al. 2005).

To understand how the differential expression of 156 genes identified by Holmes et al. (2005) relates to cell type and status, we performed Principal Component Analysis (PCA) and determined the major contributors to variation between groups. In our PCA analysis, we chose to scale and center our data. We centered because we were not interested in the relative mean of each variable. That is, the comparison of mean expression levels between different genes is not of interest to us, we only wish to know how expression levels of the same gene differ between the different groups. We also scaled our data because we are interested in fold differences and not the absolute difference in expression levels, as expression levels naturally vary between biological replicates.

From our Scree Plot (Fig. B1), we determined that the first two principal components (PCs) explain the majority of variation, with a cumulative sum of 73.4% of variation explained. The addition of subsequent PCs fails to improve the cumulative sum by a large margin (Fig. B2).

Then, we constructed a PCA plot with the first two PCs identified previously, demonstrating excellent clustering between the naive, memory, and effector T-cell types, represented by NAI, MEM, and EFFE respectively (Fig. B3). Interestingly, there is one NAI cell (blue box) grouped into the MEM group. However, after reviewing the dataset, we concluded that this may have been a labeling mistake. While all other subjects have exactly one measurement from each type of cell, subject "MEL53" appears to be missing a measure from the MEM cell type, and the outlier point was labeled as a second measurement for NAI. Hence, this point was likely a MEM cell measurement that was mislabeled as NAI.

Overall, we noticed that PC1 explained the majority of the difference between all 3 groups, while PC2 explained a majority of the difference between memory cell (MEM) and the other two cell types.

Figure B1. Scree Plot

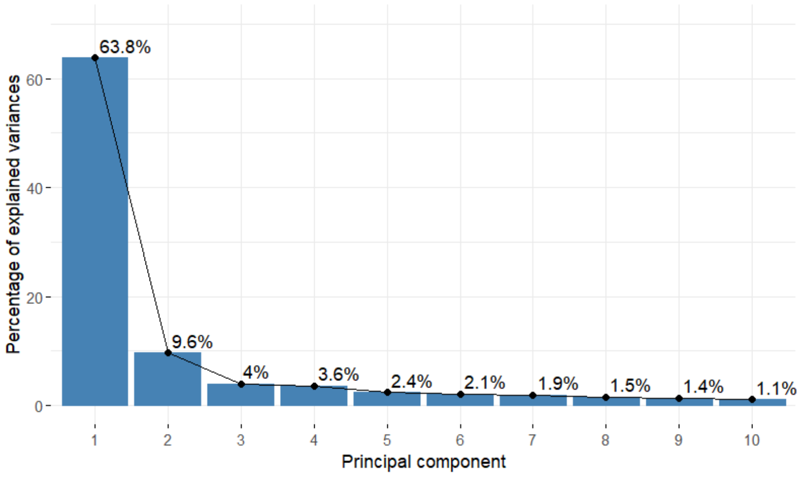
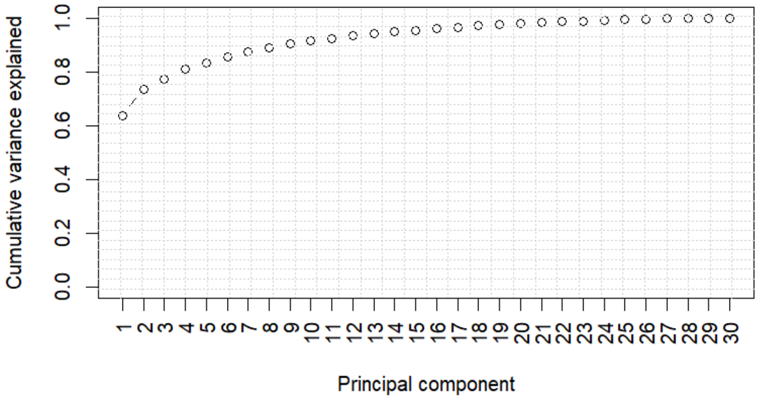


Figure B2. Cumulative sum of variance over PC



Upon examining the top 10 gene coefficients (by magnitude) in PC1 and PC2, we found that there was no overlap between the top 10 contributors to each PC. While there is a combination of contrasts that helps explain PC1 (Fig. B3), a weighted combination of genes helps explain PC2 (Fig. B4). Thus, the genes described in Fig. B3 are differentially expressed between all 3 groups, while the genes described in Fig. B4 are uniquely differentially expressed in the MEM group. This group of unique genes suggests that MEM cells are not an intermediate stage between NAI and EFFE cells, rather it represents an alternate pathway for NAI cells. Overall, this leads us to propose that NAI cells tend to move toward MEM and EFFE cells by changing the expression of top genes found in the PC1 (moving towards the left on the x-axis), however, if alternative expression of top genes described in PC2 is signaled at some intermediate stage, the cell enters an alternative path and tends towards a MEM cell rather than an EFFE cell. Moreover, we can see two points (one NAI, one EFFE) in the center area of the plot, further supporting that there is an intermediate stage between the three cell types. Overall, this helps to support the idea of parallel differentiation and an intermediate stage between the three cell types described in Holmes et al. (2005). However, our conclusions are only proposed as a theory as we cannot provide good evidence for the transition solely based on PCA. Additional analysis, such as in vitro inducement of NAI cells into EFFE and MEM cells (by artificially upregulating top genes identified in PC1 and PC2) could provide conclusive evidence of how the fate of a NAI is decided.

Figure B3. PCA biplot by cell type with top 10 explainer genes PC1

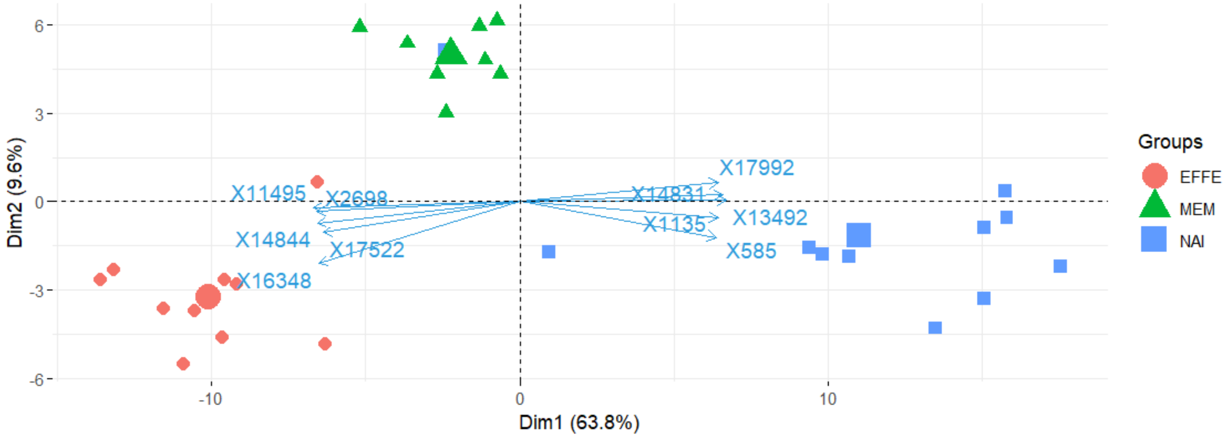
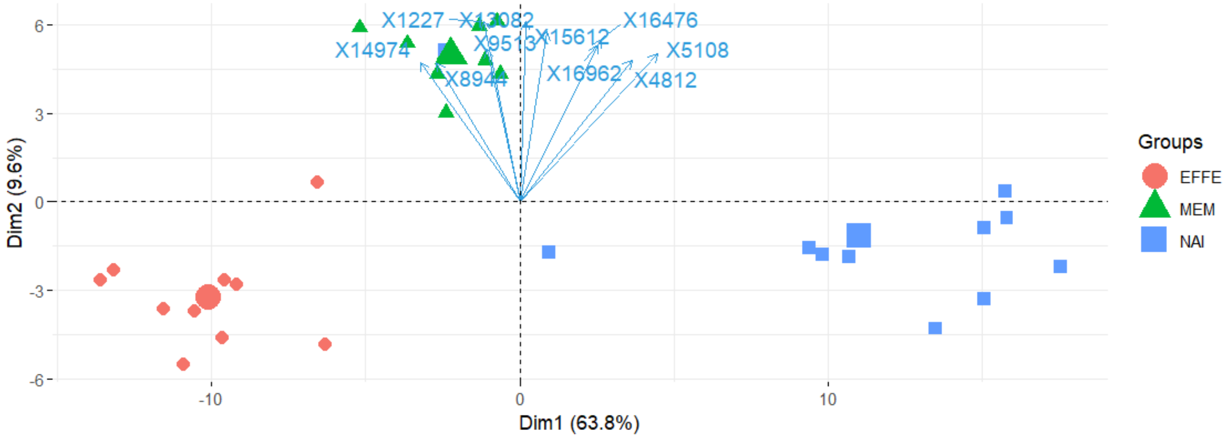


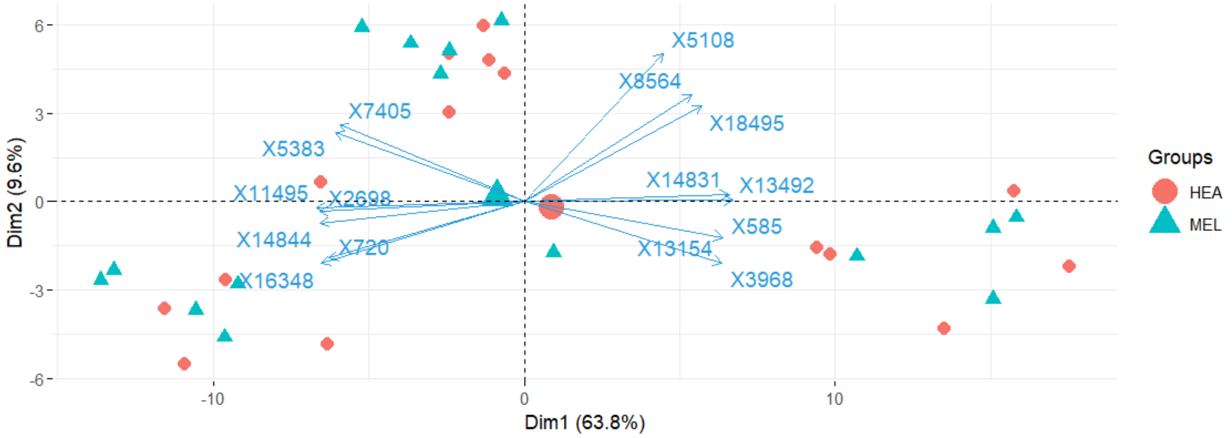
Figure B4. PCA biplot by cell type with top 10 explainer genes of PC2



Lastly, we notice that when the data is grouped by healthy or cancerous cell status, represented by HEA and MEL respectively, the two groups are very intermingled the center of the group of both groups (large triangle and circle) are close to one another, suggesting that the 158 genes in our dataset describe the variation between these two groups poorly (Fig. B5). Knowing that these 158 genes were isolated based on identified differential expression between the different cell types rather than cell status, it is not surprising that they better differentiate between different cell types rather than cancer status. However, an interesting conclusion can be made from our analysis. Xue et al. (2020) offered evidence that cancer cells can be identified by expressional differences in certain genes that promote dysregulation of the cell cycle and other systems that aid cancerous cells with proliferation and survival. Since we see here that there is no alternative expression in genes involved in differentiation between T-cell types, we can conclude that melanoma (cancerous) cells generally do not affect the regulation of T-cell differentiation and melanoma cells generally exhibit the same cell type differences as healthy T-cells. This is likely beneficial to escaping immune detection in cancer cells, as expression of these 158 genes would be similar to regular healthy cells of the same type. Holmes et al. (2005) also mirror this conclusion and state that the melanoma (cancerous) cells cannot be differentiated from healthy cells by gene expression differences that differentiate different types of T-cells.

Overall, our PCA demonstrated clear differences between gene expression in different T-cell types, identified major genes that set apart NAI, MEM, and EFFE cells, and suggested that cancerous T-cells mirror gene expression levels of healthy cells of the same type, which may be helpful for immune evasion. We also provide a good basis and gene list for future in vitro experiments where artificial inducement of differentiation may be performed to verify the role of these 158 genes.

Figure B5. PCA biplot by cell status with top 15 contributing genes



References

Blanco-Melo, D., Nilsson-Payant, B. E., Liu, W. C., Uhl, S., Hoagland, D., Møller, R., ... & Albrecht, R. A. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell, 181(5), 1036-1045.

Coomes, E. A., & Haghbayan, H. (2020). Interleukin-6 in COVID-19: a systematic review and meta-analysis. Reviews in medical virology, 30(6), 1-9.

Del Valle, D. M., Kim-Schulze, S., Huang, H. H., Beckmann, N. D., Nirenberg, S., Wang, B., ... & Gnjatic, S. (2020). An inflammatory cytokine signature predicts COVID-19 severity and survival. Nature medicine, 26(10), 1636-1643.

Giamarellos-Bourboulis, E. J., Netea, M. G., Rovina, N., Akinosoglou, K., Antoniadou, A., Antonakos, N., ... & Koutsoukou, A. (2020). Complex immune dysregulation in COVID-19 patients with severe respiratory failure. Cell host & microbe, 27(6), 992-1000.

Holmes, S., He, M., Xu, T., & Lee, P. P. (2005). Memory T cells have gene expression patterns intermediate between naïve and effector. *Proceedings of the National Academy of Sciences*, 102(15), 5519–5523.

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., ... & Cheng, Z. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395(10223), 497-506.

Lucas, C., Wong, P., Klein, J., Castro, T. B. R., Silva, J., Sundaram, M., ... & Israelow, B. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature, 584(7821), 463-469.

McElvaney, O. J., McEvoy, N. L., McElvaney, O. F., Carroll, T. P., Murphy, M. P., Dunlea, D. M., ... & Gunaratnam, C. (2018). Characterization of the inflammatory response to severe COVID-19 illness. American Journal of Respiratory and Critical Care Medicine, 202(6), 812-821.

Xue, J., Liu, Y., Wan, L., & Zhu, Y. (2020). Comprehensive analysis of differential gene expression to identify common gene signatures in multiple cancers. Medical Science Monitor, 26.

Group Work Contribution Table and Task List

Task No.	Task	Contribution by Cynthia	Contribution by Veedhi
1	Writing code - Problem 1	15%	85%
2	Writing code - Problem 2	90%	10%
3	Visualizations - Problem 1	10%	90%
4	Visualizations - Problem 2	95%	5%
5	Interpretation	50%	50%
6	Writing Report - Problem 1	20%	80%
7	Writing Report - Problem 2	80%	20%
8	Editing Report	50%	50%