Inferring Genomic Signatures in Age-related Macular Degeneration Across Different Stages

Introduction

Age-related Macular Degeneration (AMD) is a common eye disease that leads to partial or complete loss of vision. The aim of this project is to identify genes that are most responsible for causing AMD. In this project, we present a better understanding of the causal genes of AMD with statistical validation by analyzing a gene expression dataset which contained each of the four stages of the disease, with stage 1 being no disease or control group and stage 4 being the most advanced stage. We do this by discovering genes that may be responsible for causing AMD and using previous studies to conduct genomic pathway and network analysis to understand different facets of the disease.

So far, no known studies have conducted a transcriptomic analysis of AMD, making this the first of its kind. In particular, we identified the gene PMAIP1 as one of the causal AMD genes that has also been found in studies of dyslexia and is known to cause reading impairment. In the future, we hope this research is used to develop therapies to treat target genes for AMD as well as other diseases.

Background

Vision is a key sense used in nearly every aspect of life, and it relies on millions of neurons in the eye to function properly. The retina plays a large role in vision, as it is the layer where rods and cones receive waves of light and encode them as neural signals. Rods and cones are two of the six major types of neurons in the retina. Deterioration of these cells may lead to partial blindness based on the region in which the particular neuron is located. The area that we are concerned with in this study is the macula, a small and sensitive spot on the center of the retina which is necessary for central vision and spatial acuity.

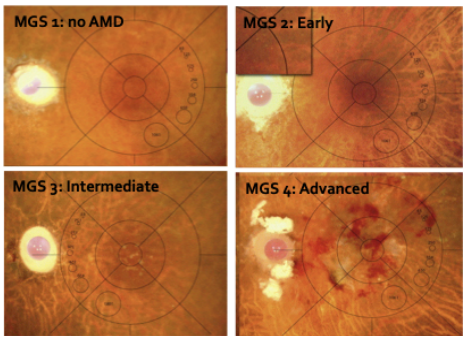

Fig.1| AMD stage progression and phenotype effects on the retina (Olsen and Feng, 2004)

AMD is a neurodegenerative disease caused by degradation of macular cells in the retina, but it generally only affects patients in its late stage--about 14% of Americans over the age 80, therefore, are subject to central vision loss as an effect (Age-related macular degeneration, 2019). This condition manifests itself as partial loss of central vision, primarily in the elderly populace. It is a complex, multifactorial disease caused by the cumulative impact of multiple genetic variants, environmental stress, and advanced aging. While there is no cure, the disease can be diagnosed at three different stages -- early, intermediate and late, as seen in Fig.1 -- with the late stage further bifurcated into “dry” and “wet” conditions. An early diagnosis can be used to manage the disease through invasive and non-invasive means. Loss of vision as an effect of AMD can be attributed to many underlying genetic mechanisms, however the specific genes and pathways have been mostly elusive to the fields of genetics and ophthalmology. Recent studies have identified genes that have a high correlation to AMD, which can be used to further isolate closely related transcriptomic data through our project.

Traditional methods include transcriptome analysis, which takes RNA-seq data and creates an expression profile from a dense sequencing process. These types of analyses formulate expression quantitative trait loci (eQTL) that might be useful in predicting variation in levels of RNA expression across experimental groups. One study in particular, conducted by Ratnapriya et al, has identified a total of 37 loci (including 34 from earlier Genome-wide association studies) using transcriptome analysis (Ratnapriya, 2018).

Differential expression is a common method to quantify changes in the way that certain genes are expressed between groups (e.g. disease, non-disease) in a study. For most genetic-based diseases, differential expression has been useful for isolating a handful of genes and was considered to be a potential solution for AMD (Morgan et al, 2014). However, when applied in study, this method was not helpful in selecting key genes: only as a batch predictor (Ratnapriya, 2018). Other gene targeting methods for diseases similar to AMD include dimensionality reduction and feature selection on the RNA-seq counts matrices. Singular value decomposition (SVD) is a common data science method for decomposing linear systems into their fundamental components, but it is also useful for removing redundant genes (i.e genes that

have low variance across the experimental groups) and therefore reducing the complexity of the gene filtering problem (Alter et al, 2000). Feature selection methods, like Lasso and logistic regression, are also useful for reducing the number of genes from their original expression counts matrices--which are usually very high complexity. A study on the applications of feature selection for microarray analysis has deemed group Lasso with clustering as a strong classifier (Ma et al, 2007).

Literature has also provided us with ways to understand how we can understand the biology behind the genes identified using statistical methods. Common avenues to understand the interactions of genes and they cause specific diseases are through pathway and network analysis. A pathway is a linked series of actions among genes in a cell that lead to a certain product or a change. Since genes do not work alone, it is useful to understand how other known genes interact with the genes found to produce the effect that eventually causes the disease. In a similar vein, networks are a group of known proteomic or genomic pathways that act together. Modeling a given pathway in networks provides insight into the biology of the findings (Khatri et al, 2012). Moreover, it adds scientific validity to our results and can guide medical decision making around drug targets for AMD.

This project is significant for multiple reasons. First, AMD and similar diseases do not have much concrete support for determining genes associated with distinct stages of the disease. Being able to detect genetic predisposition, the affected genes and networks, at an early stage of a disease is crucial to further prevention and treatment (Tuo et al, 2004). Second, genomics is a highly data-driven field, and therefore the results of our project may act as a proof of concept for creating machine learning models to tackle high-throughput biological data, specifically RNA-seq. Our overall goal as data scientists is to generate an analytical pipeline that will address the challenges of subtle gene expression changes associated with AMD, such that they can also be easily applied to similar diseases.

Through this project, we hope to identify transcriptomic expressions of AMD using a genomic dataset collected from the retinas of 500 donors at the National Eye Institute, Bethesda. Finally, our analysis of the dataset to generate gene expressions will be fed into existing methods such as GO enrichment analysis to perform biological network analysis.

Objectives

The long term goal of the project is to bridge the gap between genomic signatures and biological pathways that eventually cause AMD. We have three main objectives:

1. Explore genes and genomic pathways associated with AMD stages.
2. Discover genes and pathways predictive of distinct AMD stages.
3. Explore genomic network signatures of AMD and specific changes in networks associated with the AMD stages.

Data Science Pipeline

For the complete DS pipeline, see Appendix.

Data Description & Visualization

We were **initially** given three datasets:

1. Meta retina dataset, which contains donor's information such as donor id, rnaseq id, disease stage with 4 stages, the mgs_level (Minnesota Grading System-MGS, MGS1 donor retinas demonstrated no AMD features and served as controls, whereas MGS2 to MGS4 samples represented progressively more severe disease stages), genotype id, age, sex, race, mgs level, cause of death, etc. These attributes provide us with information about how the test was conducted at the National Eye Institute.

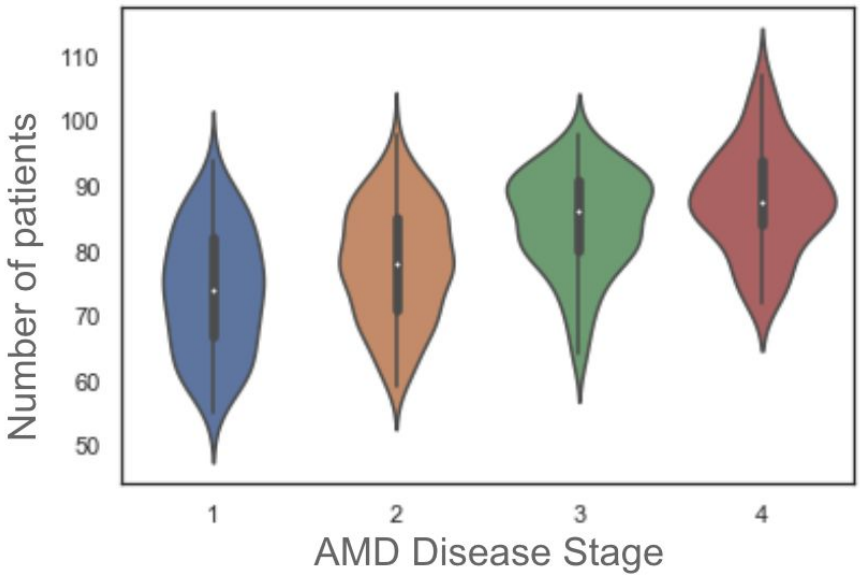

Fig.2| a) Violin plots of the relationship between the AMD disease stage and samples' age.

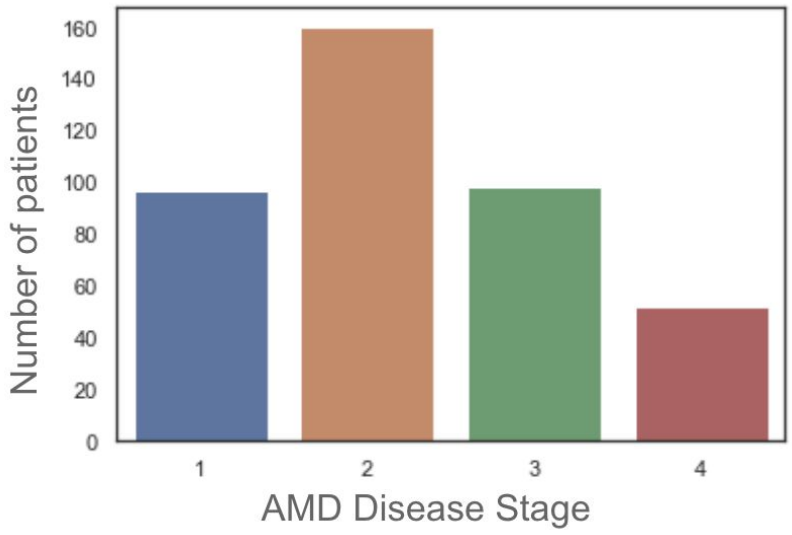

Fig.2| b) Barplots of the number of samples at each AMD disease stage.

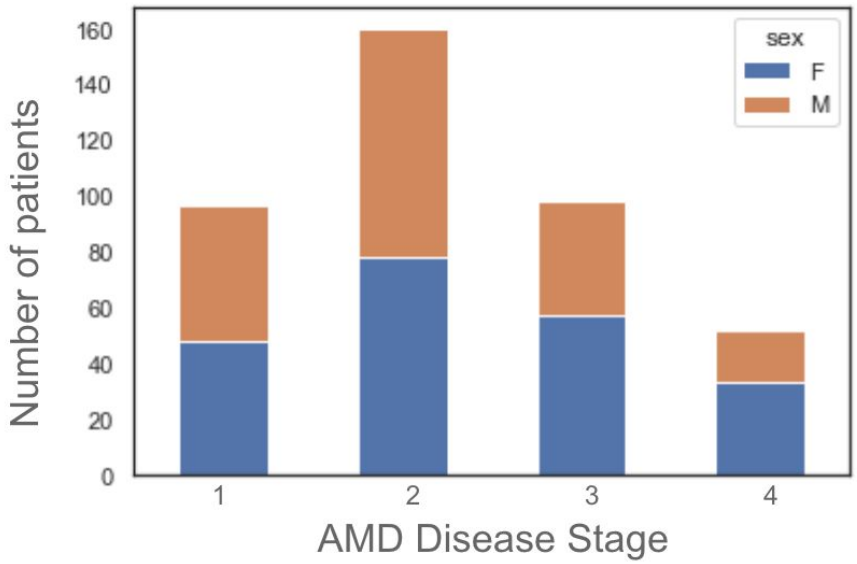

Fig.2| c) Barplots of the patients gender proportion for each stage.

2. Matrix of mRNA counts for every gene in each patient contains the most vital, raw information for gene counts. We plan to use this dataset for our project.

3. Matrix of batch-corrected mRNA counts(log2 counts per mil reads normalized) for every gene in each patient. This dataset is equivalent to the previous one, but is normalized and batch-corrected for non-biological factors.

We first used dimension reduction methods like t-SNE and UMAP to visualize the data and found that t-SNE map showed two big clusters(Fig.3 a). Then we found that the batch might still affect the dataset, according to meta retina dataset, we noticed that features like sex, rna_isolation_batch and library_prepper might cause the abnormality. Therefore, the batch corrected data was plotted again and labeled by these three factors(Fig.3 b-d). Shown by these figures, we conclude that the gender batch still affects the whole dataset. By removing the XY chromosome genes we found the cluster disappears (Fig.3 e). We realized that the matrix of batch corrected dataset was not useful for performing further data analysis, since the automated scripts to batch correct data were very opaque.

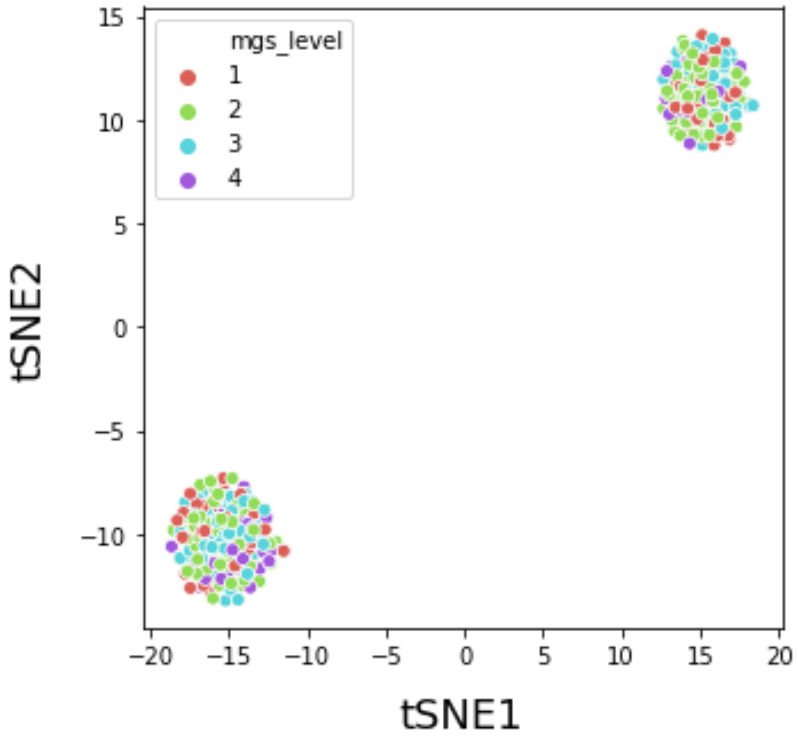

Fig.3| a) t-SNE plots for batched corrected dataset and labeled by AMD stage(mgs_level).

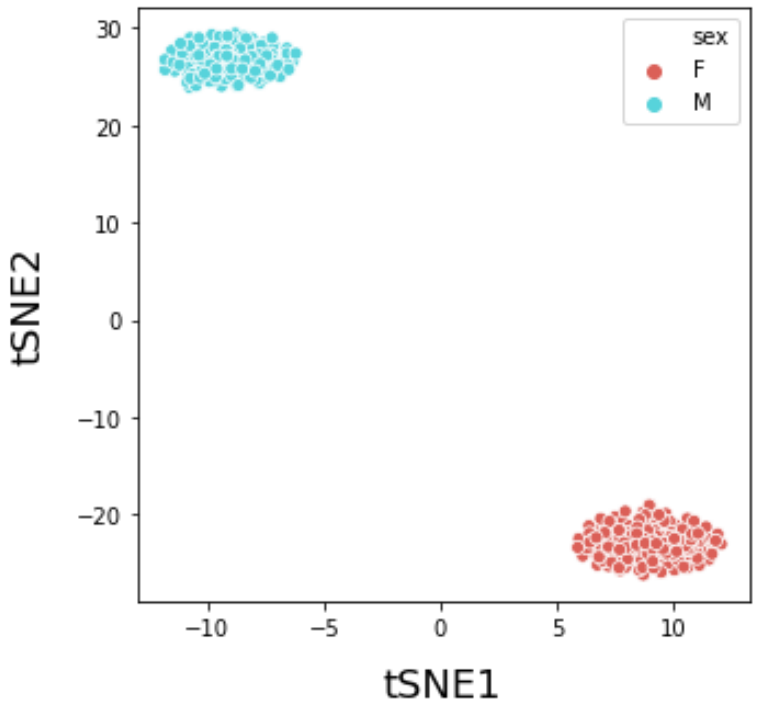

Fig.3| b) t-SNE plots for batched corrected dataset and labeled by gender.

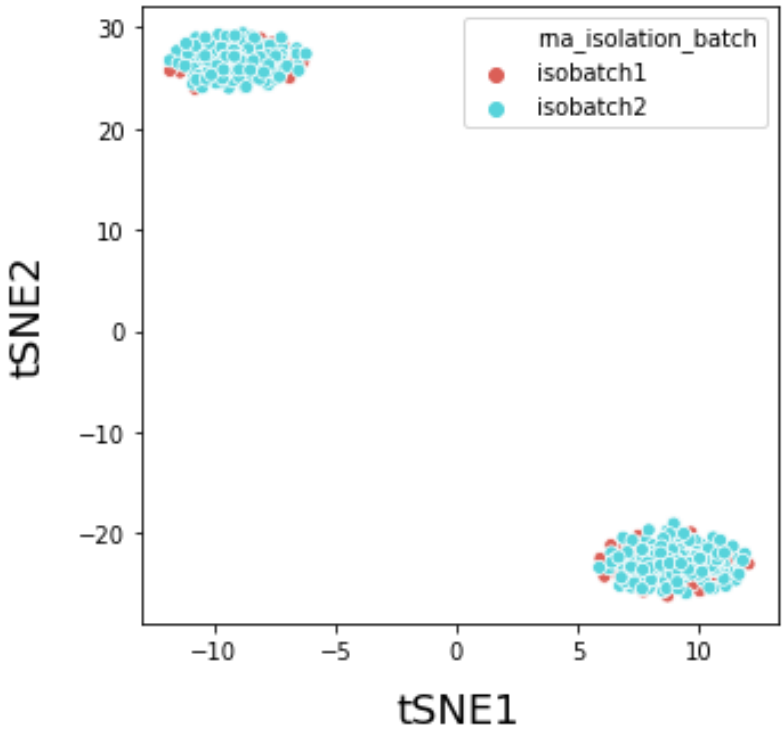

Fig.3| c) t-SNE plots for batched corrected dataset and labeled by rna_isolation_batch.

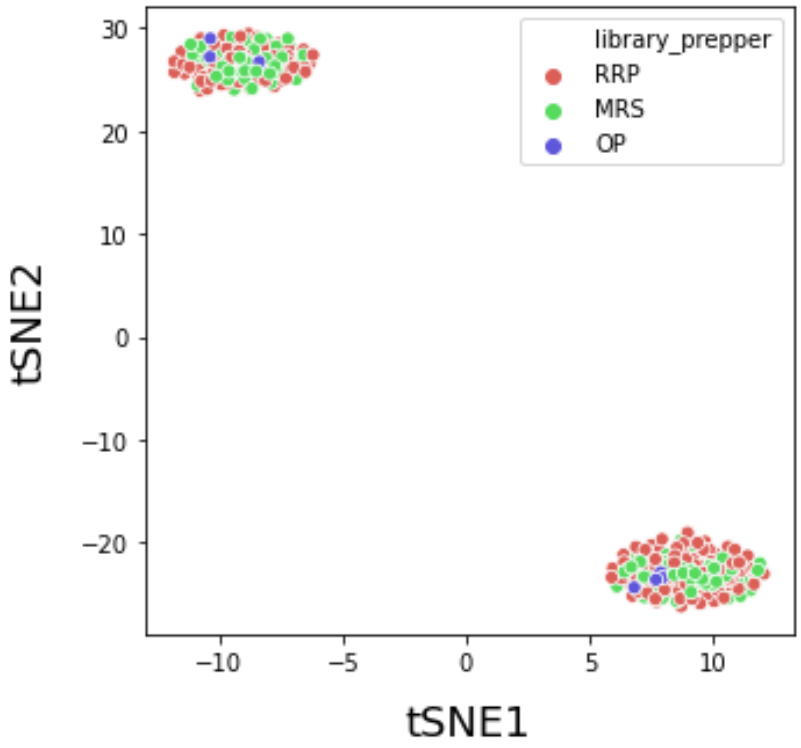

Fig.3| d) t-SNE plots for batched corrected dataset and labeled by library_prepper.

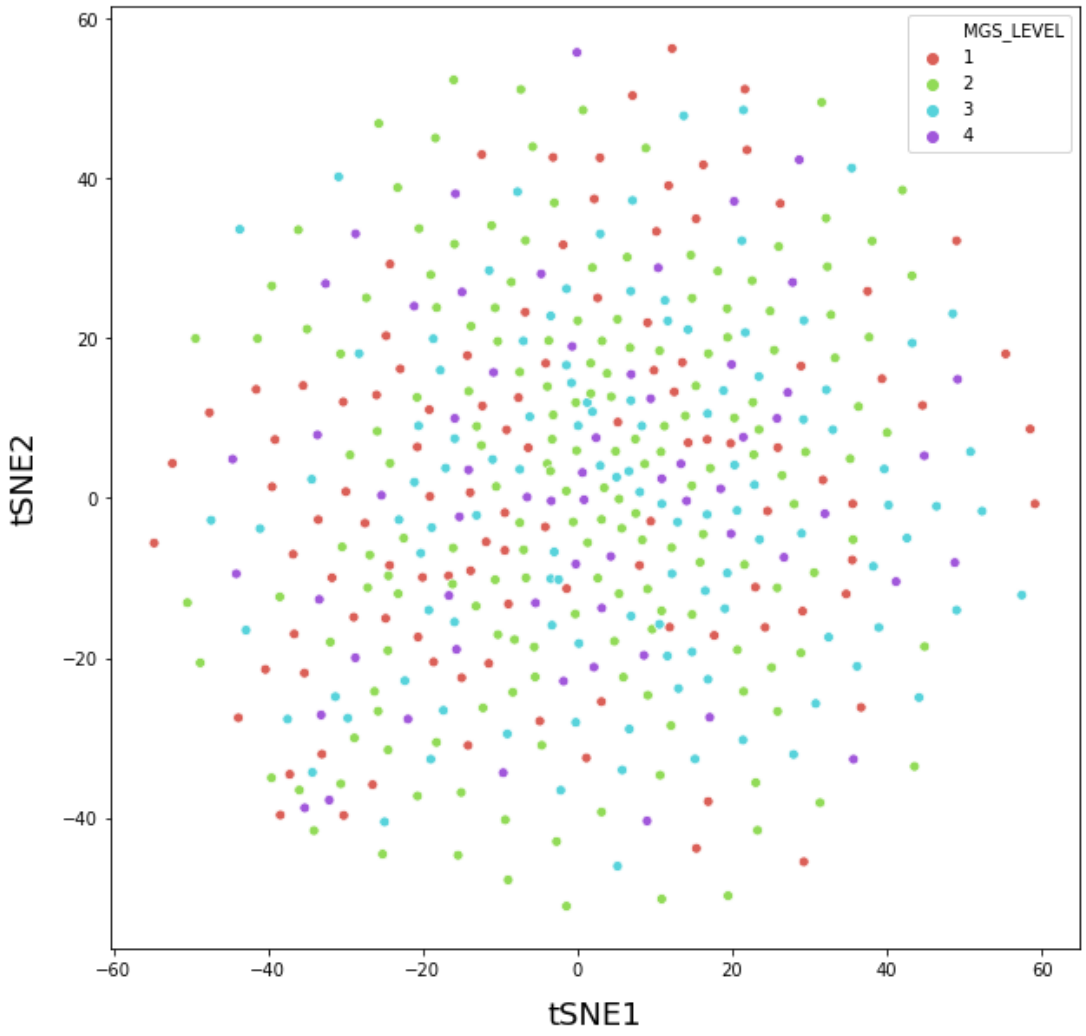

Fig.3| e) removing X&Y chromosomes for batched corrected dataset and labeled AMD disease stage.

Instead, we decided to use the original mRNA counts dataset, which contained about 58,051 genes from 453 patients with 4 stages, and a dataset mapping each geneID to its common gene name which was provided by our sponsor. We joined the two datasets on column "gene_name" to generate a new dataset with 58,051 genes and their common names as well as their gene IDs. Before splitting the dataset into train and test, we first check to make sure that gender batch (the x,y chromosome) or other batches do not have an effect on the dataset (Fig3.f). We then split this new dataset into a 90:10 ratio by number of patients of training data (train_data.csv) and test data (test_data.csv).

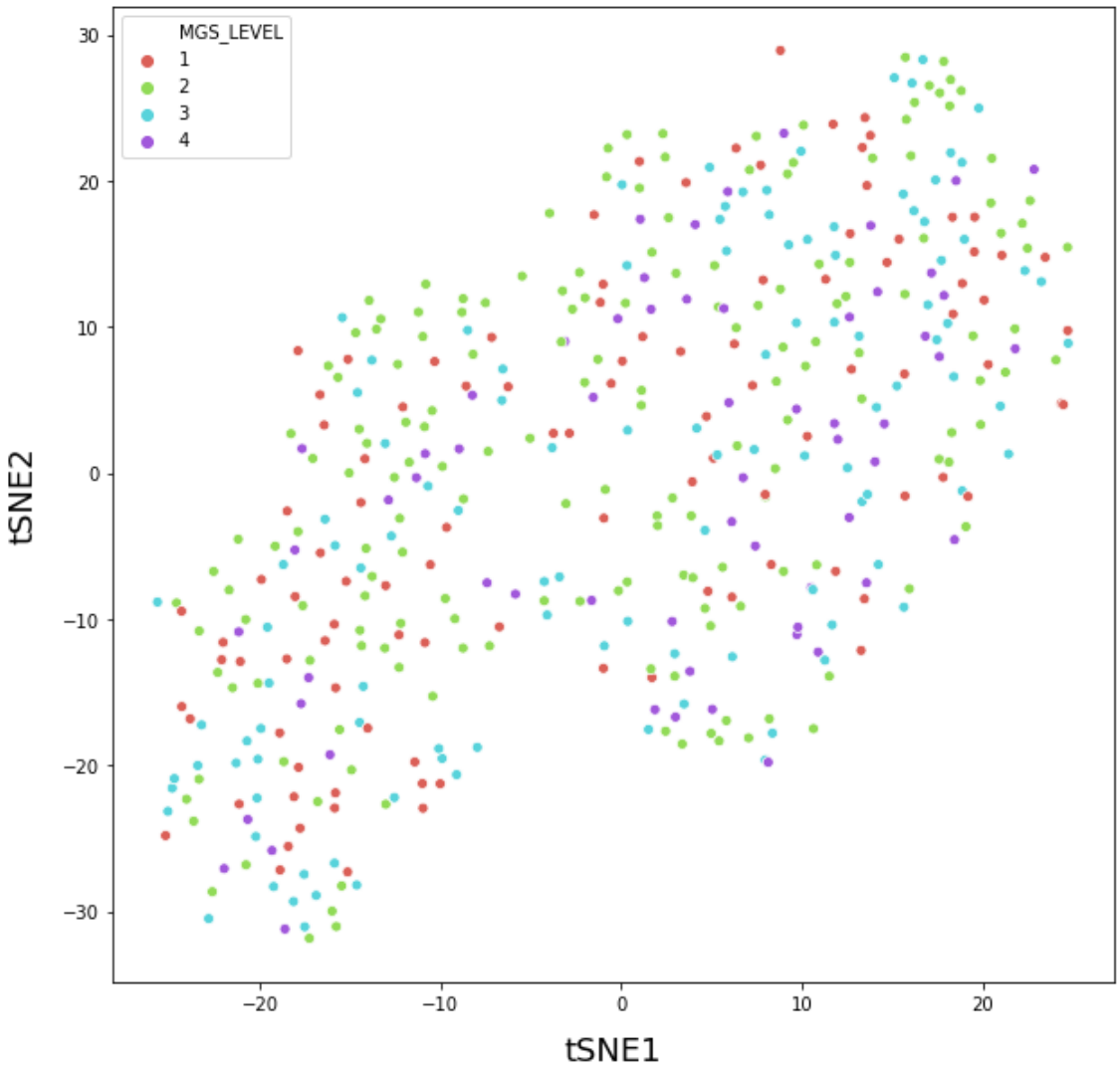

Fig.3|f) t-SNE plots for original mRna counts labeled by AMD stage (mgs_level).

In addition, in order to have a clear understanding of how batch effects affect the gene expressions, we draw many violin plots based on the expression of a specific gene and several factors of great importance (age, gender, library Prepper, RNA isolation batch, etc.) to check the distributions of gene expressions among different groups and how they are distinct from each other. Fig.4 shows two examples of violin plots based on gene ENSG00000000419.

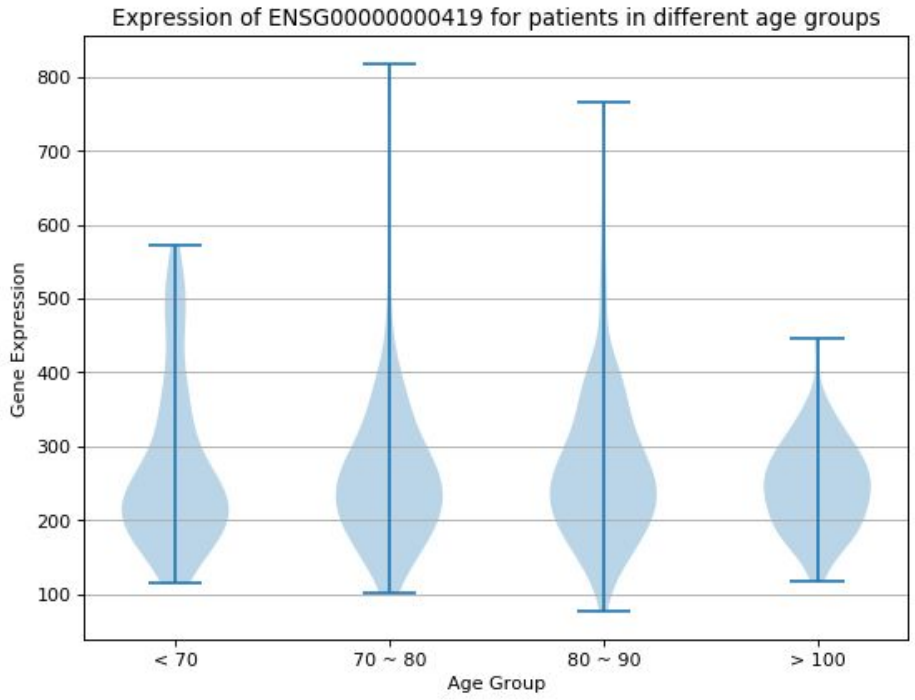

Fig.4|a Violin Plot that describes the expression of a specific gene for different age groups

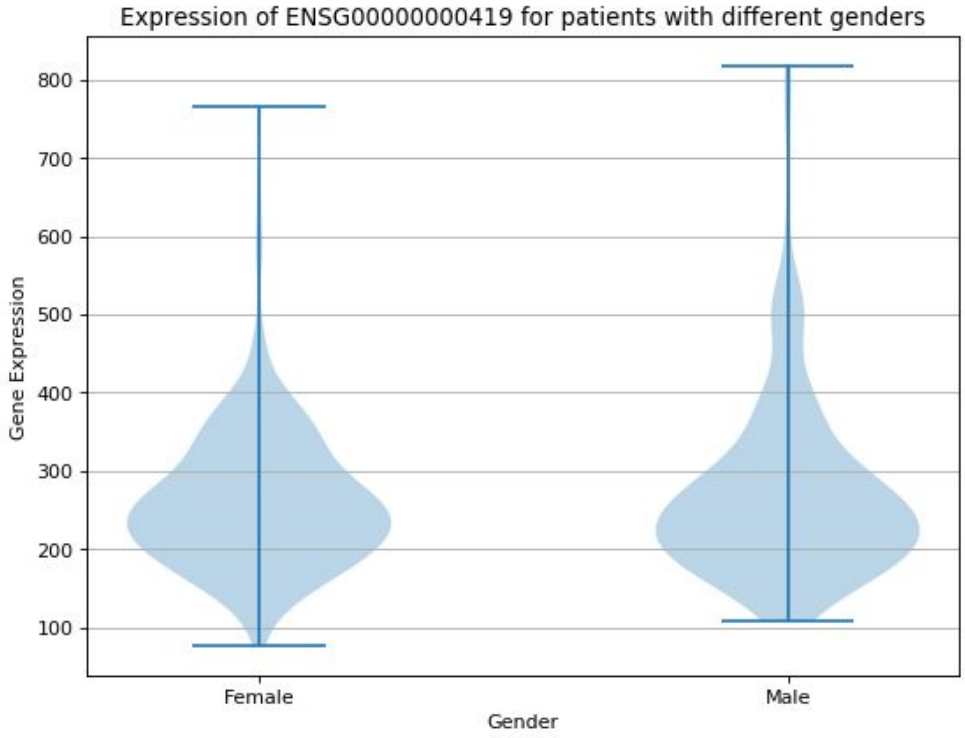

Fig.4|b Violin Plot that describes the expression of a specific gene for different gender

Data Wrangling

Normalization and Low Variance filtering

Due to a large number of genes in the raw dataset, we needed to identify the problematic genes among 58,051 genes. Therefore, specific normalization methods and dimension reduction methods were needed.

We researched well known normalization methods. Some of them, in particular, RPKM, not only remove the gene library depth effect but also consider the gene length's effect. However, we realized that we only needed to remove the influence of gene library depth. Therefore, we decided to normalize using our own method. For each gene, we divided it by patient's library depth, which is the maximum of all genes for each patient. Then we multiplied all gene expression levels by the maximum library depth among all patients. The normalization ensured that the mRNA counts for each gene were numerically scaled to a similar value, removing any bias that it may have caused. Finally, we applied log2(1+counts) normalization across patients to reduce the scale of the numbers.

We then plotted to see what the variances of genes in our dataset after normalization was. The histogram shows that most genes have very low variance (Fig.5). This meant that we can significantly reduce our features using low variance filtering. Therefore, we removed genes with a variance threshold of less than 0.05 log2 count units. We further compared our resulting dataset which contained about 18,000 genes by comparing with other known datasets (batch corrected dataset and normalized dataset) we received from our sponsor. We found that the overlap between the genes obtained by our method and the given data sets is about 98%. This again validated that our data pre-processing was scientifically grounded. Our sponsor also checked that the genes resulting from differential expression were all in our filtered dataset. However, the number of features still remained too large for effective modeling techniques. Hence, we decided to explore other feature selection methods.

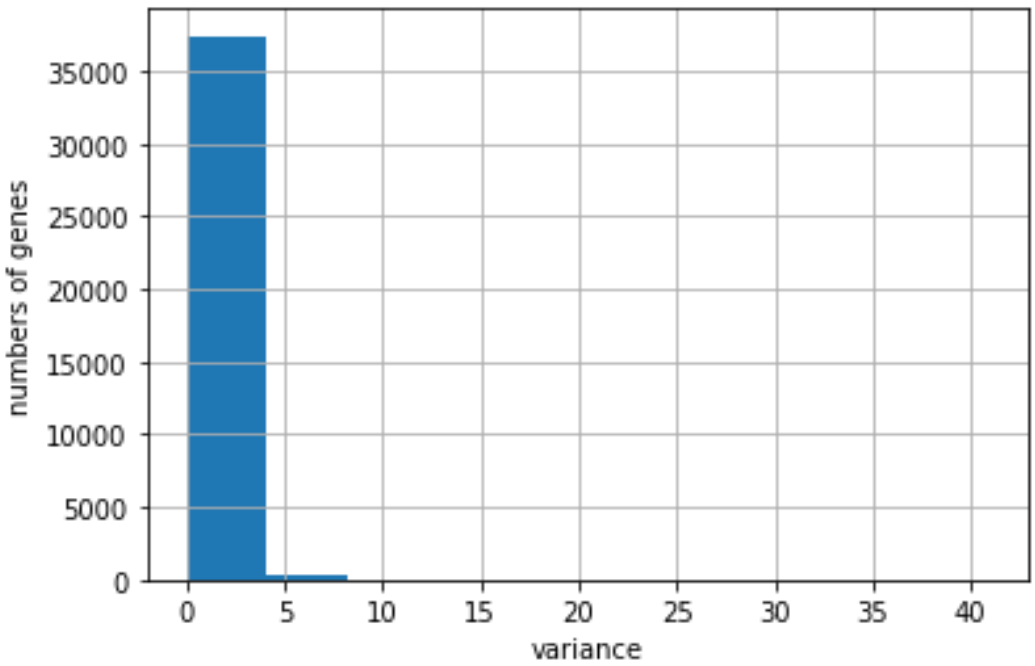

Fig.5| Histograms for low variance filtering

ANOVA

We discussed with our sponsor and confirmed that the ~37 k features selected through low variance filtering was sufficient to model from, however, some models may overfit on some genes at this stage that have high variance in only one or two stages. To further reduce the dimensions of the dataset, we used one-way ANOVA (Analysis of Variance) on all four groups in a gene-wise manner. This test was performed using SciPy package's stats.f_oneway function in Python, and we compared the difference of means of the genes expression counts. Our null hypothesis was that there are no differences between the gene expression of a particular gene across all four groups. After performing ANOVA, we found statistically significant differences ($p < 0.05$) in 4,480 out of ~37,000 resulting genes after low variance filtering. Later, we used a clustered heatmap to compare genes of all four groups. Fig.6.a illustrates the resulting clustered heatmap. It was hard to distinguish differences among genes since the differences were small and the number of genes were too many to visualize. But several clusters show that after low variance and ANOVA analysis some genes have stronger relation to some stages (Fig.6.b). The presence of vertical clusters especially indicates that genes share a similar expression (Fig.6.a).

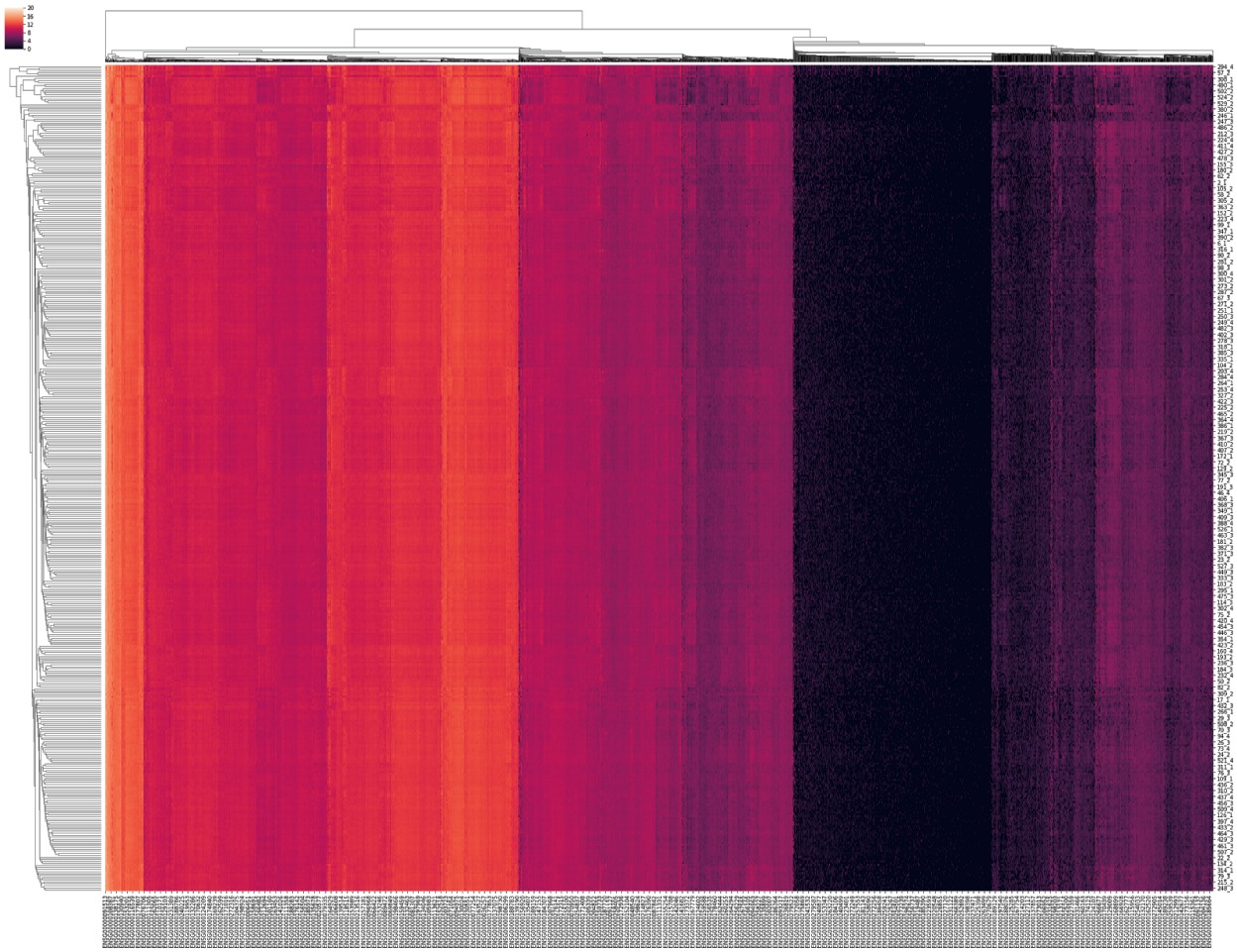

Fig.6| a) Colorbar of all 4438 genes on the y-axis and patients on the x-axis. The number before "-" is the sample's ID and the number after "-" is the AMD disease stage.

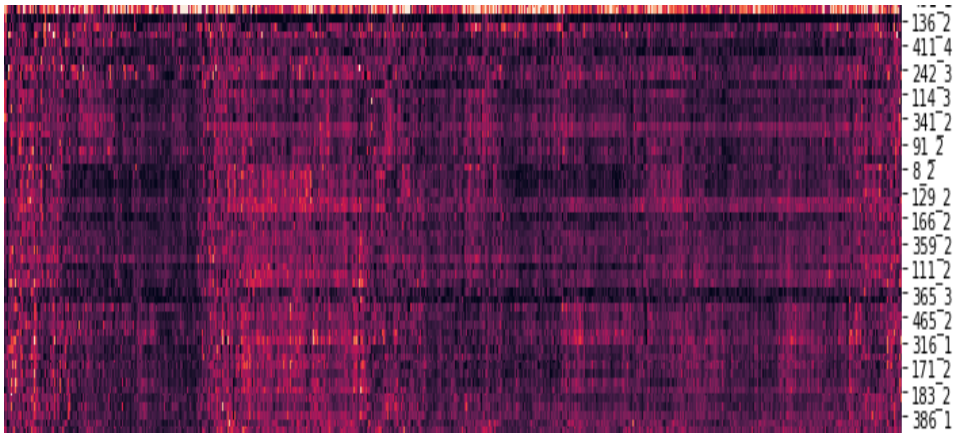

Fig.6| b) Part of Cluster Heatmap shows some genes are more related to stage 2 and 3. The number before "-" is the sample's ID and the number after "-" is the AMD disease stage.

Method and Models

The 4438 genes obtained after low variance filtering and ANOVA analysis is the dataset we will be using to train models to find the most significant genes across different stages using different methods. Specifically, we used Multinomial Regression with Lasso Penalty, Random Forest and XGBoost classifiers to find predictor genes of all 4 stages. In our Multinomial Regression model, we applied an L1 penalty to directly compare control samples with patients in each stage. Both tree-based models were run both with and without the SMOTE oversampling technique, in response to the class imbalances of our dataset.

In each model, we only touched the 90% training dataset. Within the training dataset, we again split into 90% training and 10% validation dataset so that we could train our models using 10 fold cross fold validation. Our goal was to correctly classify all 4 stages. In these classification models, the target variables were all 4 stages. Our input was the dataset obtained after low variance filtering and ANOVA.

Multinomial Regression with L1 penalty

At first, we tried different feature selection methods to find the best model for a multi-classification problem. In this model, the tuning parameter lambda is chosen by cross validation, and the model chooses the lambda where multinomial deviance is minimized.

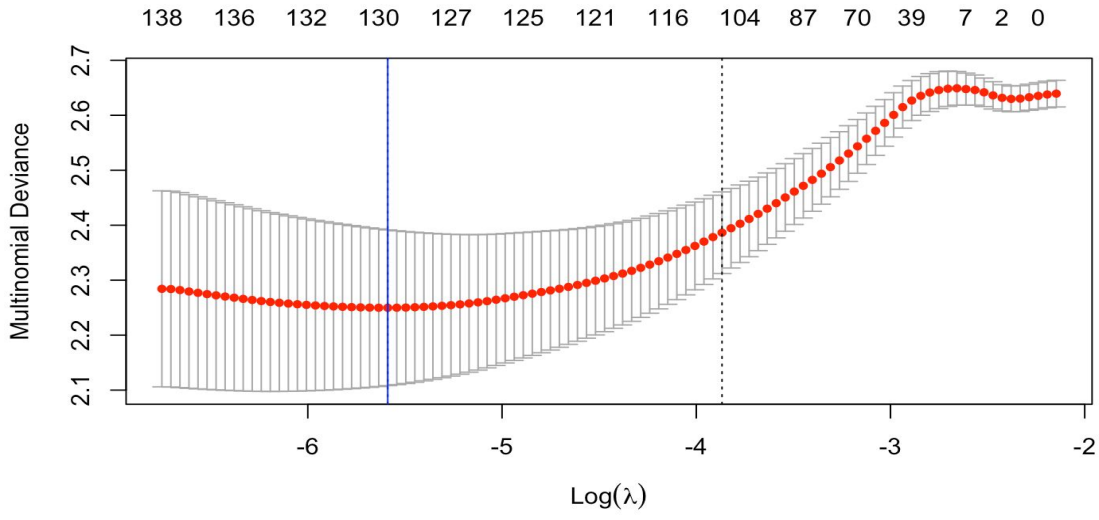

Fig.7| From 10-fold cross validation, the model chooses lambda that minimizes multinomial deviance

We first used multinomial regression with lasso. Through cross validation, the model selects genes that are important for each stage of the disease. Genes that appear at least 5 times in 10 fold cross-validation are further selected to eliminate genes that are chosen by chance. In summary, 304 genes for stage 1, 424 genes for stage 2, 249 genes for stage 3, and 249 genes for stage 4 are selected. We were surprised to see that 66 genes overlap in stage 1 and stage 2, and 20 genes overlap between stage 1 and stage 4. This is reasonable because stage 1 and stage 2 are similar in terms of disease progression, whereas the difference between stage 1 and stage 4 is more significant.

We also looked at binary classification to identify important genes by comparing the control with the group in each stage. In stage 1 vs stage 2 setup, our model found 177 significant genes. For stage 1 vs stage3 and stage 1 vs stage 4 setup, the model chose 136 and 91 genes respectively. We also looked at if there is any overlap of genes found. Among all these regressions, we found 9 genes that occur in all stage-comparisons. Among these genes include ATP2C2. Diseases associated with ATP2C2 include Dyslexia, which is a brain-based type of learning disability that impairs a person's ability to read. Furthermore, as shown in the correlation matrix heatmap in the Fig.8.a, Fig.8.b and Fig.8.c, significant genes found in each comparison do not have strong correlations to each other, and this suggests that genes selected by the model in each comparisons are unique and independent to each other.

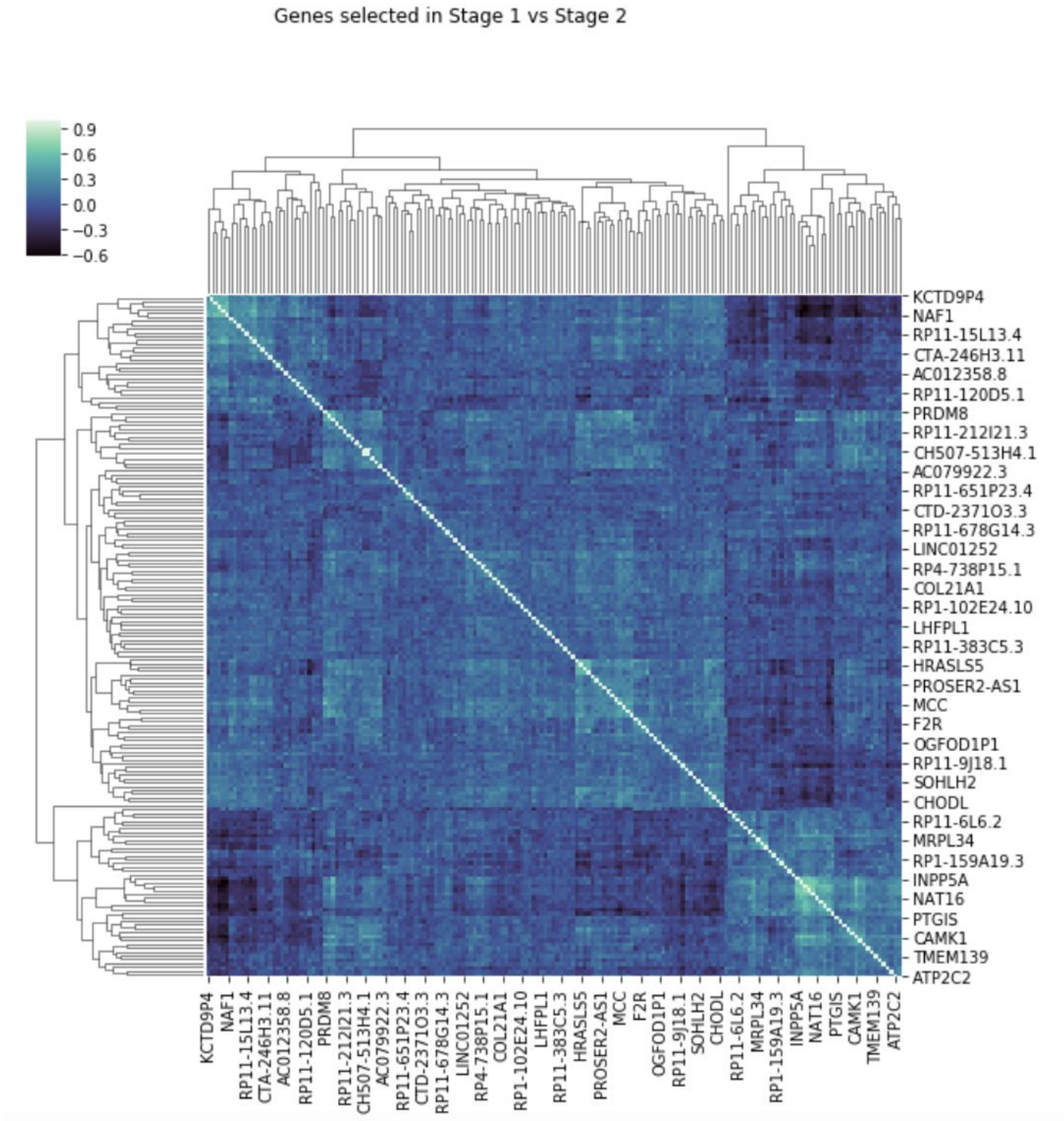

Fig.8|a) Correlation matrix of between genes selected by the model when comparing control group with each stage of disease Stages 1 vs 2

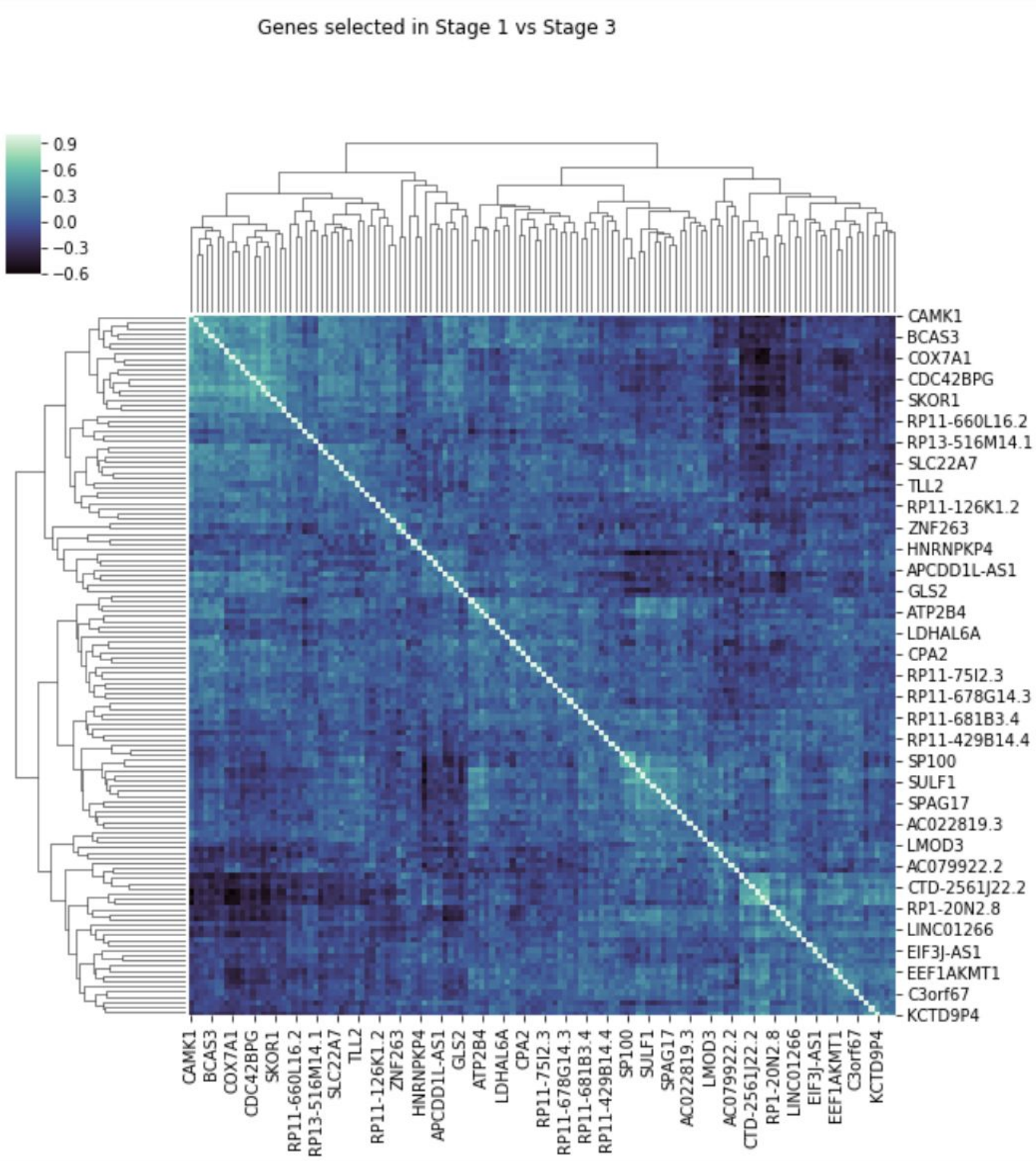

Fig.8| b) Correlation matrix of between genes selected by the model when comparing control group with each stage of disease stages 1 vs 3

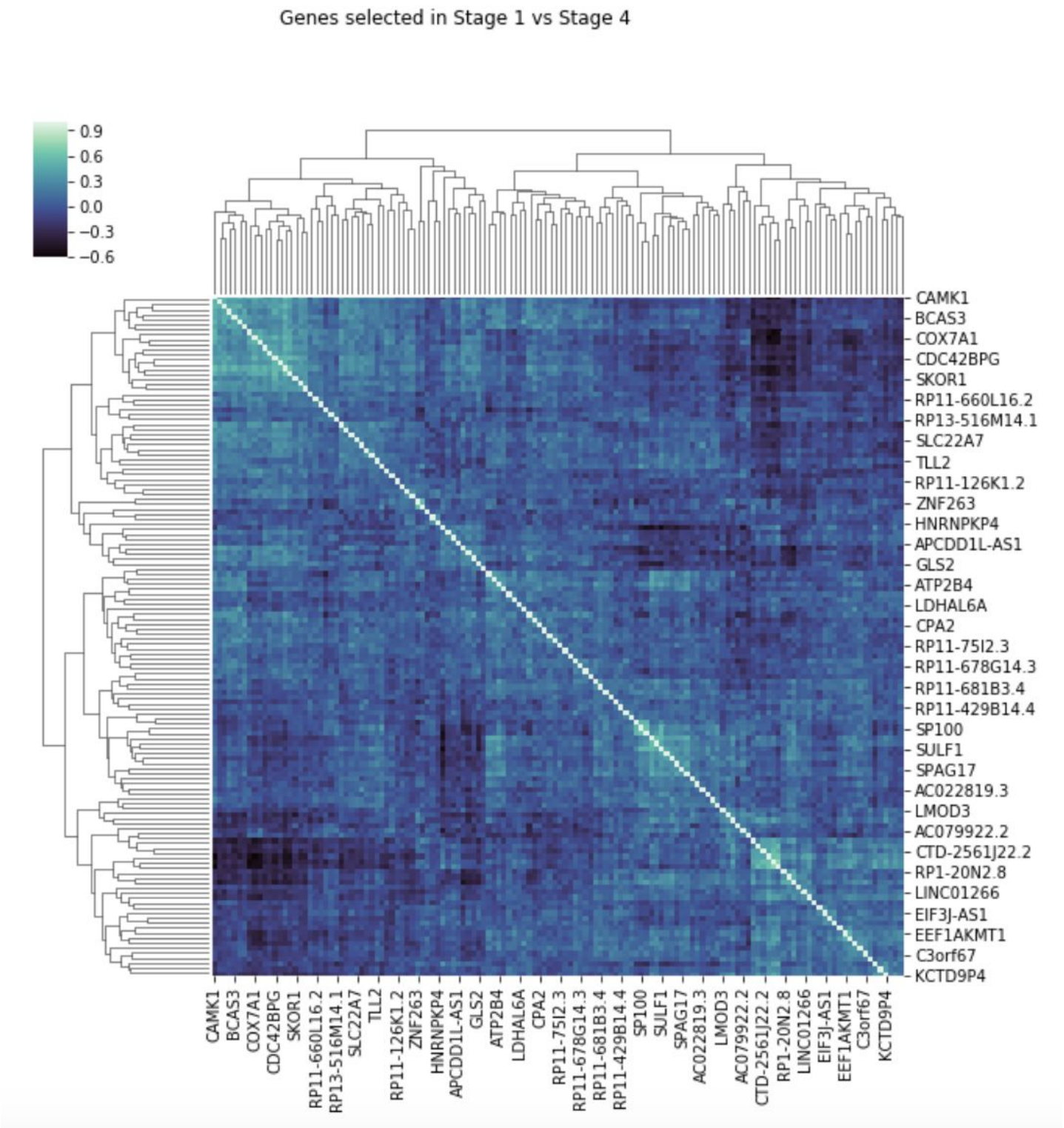

Fig.8|c) Correlation matrix of between genes selected by the model when comparing control group with each stage of disease stages 1 vs 4

Random Forest

The Random Forest model ranks the significance of all filtered genes when we use their degrees of expression as input to predict which AMD stage each patient is in. In order to improve the performance of the model, we compare results given by a normal Random Forest classifier and a balanced one, introduce synthetic samples with SMOTE, and tune main hyperparameters, including the number of trees in forest and maximum depth of a tree. Finally, the normal Random Forest Classifier trained with the data preprocessed by SMOTE achieved an accuracy score around 42%. With the ranking list given by this model, we can reasonably determine our desired level of significance when choosing genes for pathway analysis. Fig.9 shows how we tune hyperparameters and compare performances of models in various cases, where each color represents a distinct type of model (Red: Random Forest Classifier (RF); Yellow: Balanced Random Forest Classifier (BRF); Blue: RF with SMOTE; Green: BRF with SMOTE).

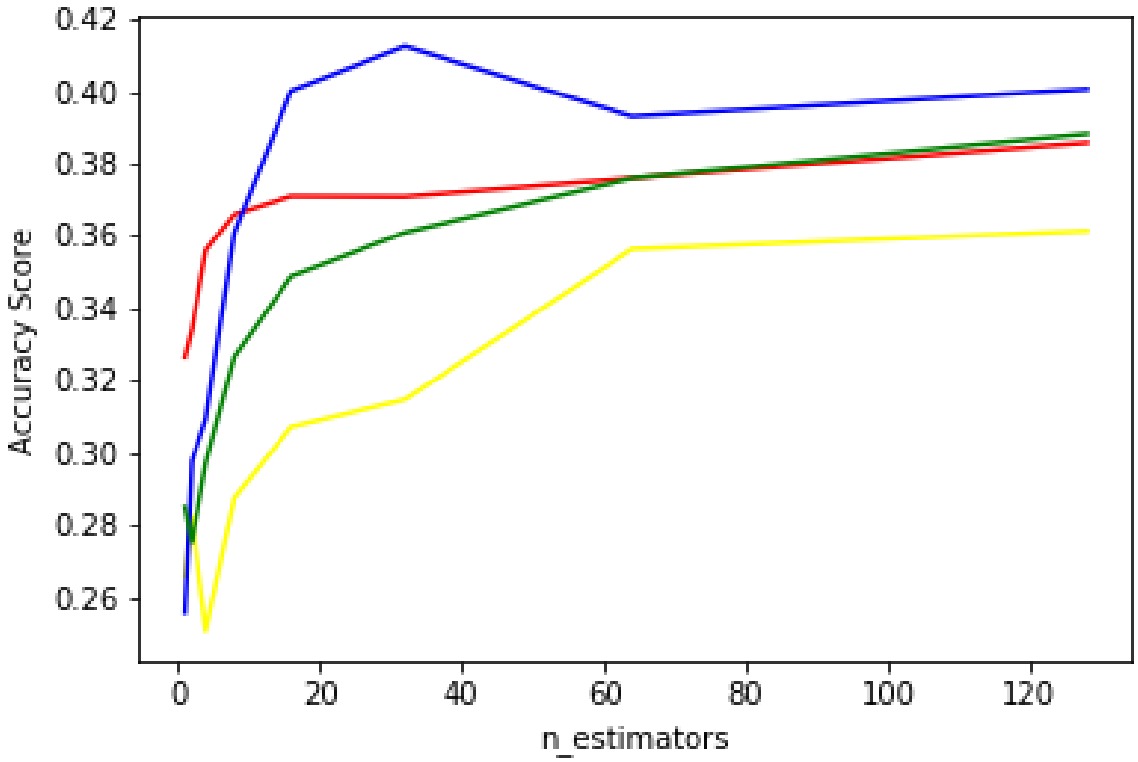

Fig.9|a Tuning the number of trees in Random Forest.

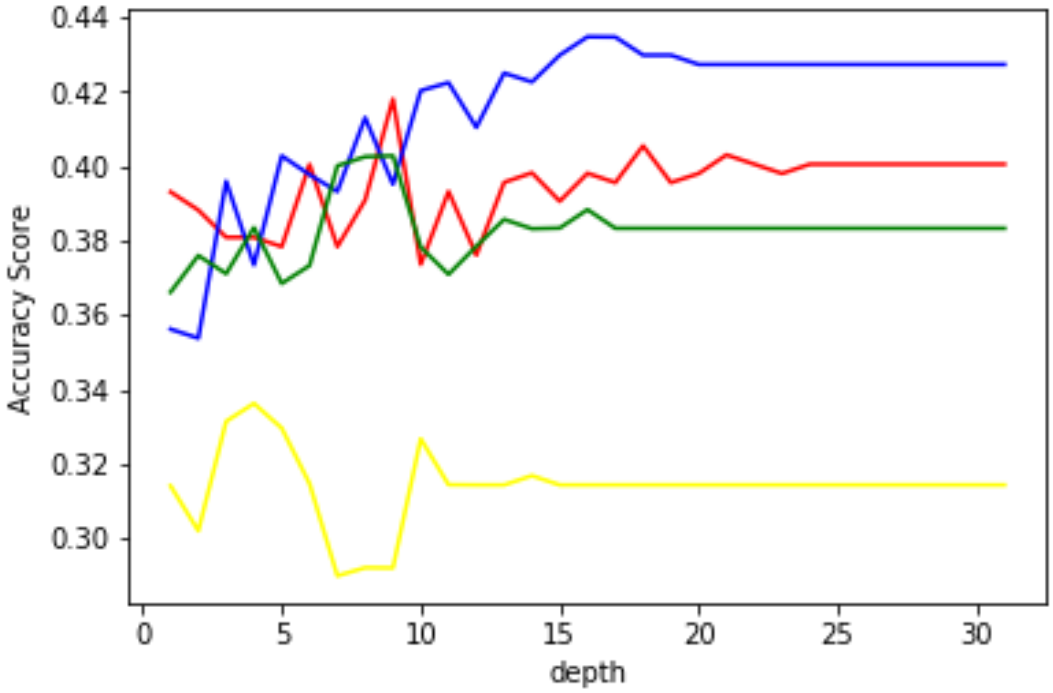

Fig.9|b Tuning the maximum depth of a tree in Random Forest

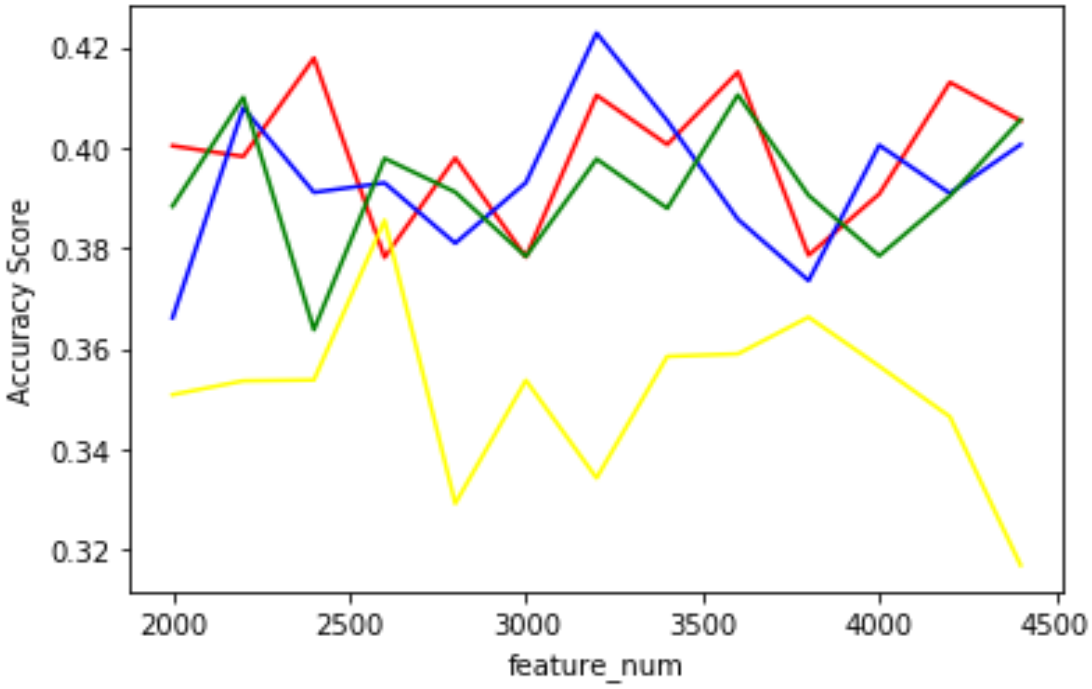

Fig.9|c Tuning the maximum number of features (gene expressions) to be considered once

XGBoost

The XGBoost model offers perks very similar to that of Random Forest in that it is a boosted decision tree algorithm, but we can use Python's XGB package to tune parameters for optimal results. Ranking genes by importance using the classifier is also enabled by Python XGBoost API. Using a multinomial softmax classifier, it is possible to determine the stage of a given patient based on their gene expression of the selected (filtered) genes, and we can easily validate the model after having split our train and test data early on in data wrangling. We trained our model on 90% of patients (with 10-fold cross validation) in our filtered gene expression data matrix, with labels being the stage of AMD for each patient.

Using grid search hyperparameter tuning, as seen in Fig.10 below, we selected the optimal inputs to the XGBoost model and achieved an accuracy score of 50.5% without SMOTE and 51.5% with SMOTE. This score was expected due to the multiclass aspect of our data; in a binary classifier with the same parameters we achieved an 85% accuracy score between classes 1 and 4, the classes expected to be most separable. Since our classes were not extremely imbalanced--no class took the majority of the dataset, and the minority class still covered about an eighth of the dataset--it made sense that the SMOTE was not effective in improving our results. Each of our XGBoost models are able to output a gene list ranked by importance in the model, as well as a classifier capable of interpreting the test data or other AMD data sets.

Hyperparameter name	**Grid Search values**	**Grid Search values w/ SMOTE**
Max Tree Depth	1, 3, **5**	1, 3, **5**
Minimum Child Weight	1, **3**, 5	1, **3**, 5
L1 Regularization α	1×10^{-5}, **0.1**, 1	1×10^{-5}, **0.1**, 1
Learning Rate	0.01, **0.1**	0.01, **0.1**
Subsample Ratio	0.6, 0.7, **0.8**	0.6, 0.7, **0.8**
Subsample Ratio of Columns	**0.6**, 0.7, 0.8	**0.6**, 0.7, 0.8
Number of Estimators	**100**, 300	**100**, 300
Sampling Strategy	N/A	Minority, **Not Majority**

Fig.10| Table of hyperparameters for tuning the XGBoost model. The parameters that achieved the best cross-validation score (altogether, not individually) are highlighted in red.

Results

We cross-checked the important genes that are chosen by all three models. And our models identified genes MOXD1 and PMAIP1, which genes are verified to be differentially expressed genes. These genes have potential to be closely related to AMD disease, as they show difference in gene expression level as the disease progresses.

For stability testing, we did bootstrapping of 500 iterations to make sure that genes selected repeatedly are not by chance. As shown in the Fig.11, the top 10 genes selected appear repeatedly in 500 iterations.

bootstrap_stage1_gen

	s1_var_names	Freq
123	RP3.395M20.9	399
903	RN7SL505P	378
357	CTD.2376I4.2	346
800	OSTCP8	343
326	CTB.134H23.2	340
9	AC002984.2	339
919	RNU6.548P	337
350	CTD.2161E19.1	336
324	CTB.108O6.2	311
126	RP5.1009E24.8	289

bootstrap_stage2_gen

	s2_var_names	Freq
65	AC068831.6	438
202	bP.2171C21.2	417
110	RP11.218C14.5	392
117	RP11.370A5.2	378
151	SMIM2	378
411	CTD.3116E22.7	358
130	RP11.681L8.1	343
680	KRT18P58	324
401	CTD.2555O16.1	320
119	RP11.408H1.3	312

bootstrap_stage3_gen

	s3_var_names	Freq
438	FLT1P1	399
806	RBPJP2	375
539	JSRP1	365
956	RP11.326E7.1	361
555	KRR1P1	358
67	AC104304.4	354
871	RP11.118E18.4	350
131	SSU72P2	345
492	HIST1H3G	344
860	RP11.109P14.2	336

bootstrap_stage4_gene

	s4_var_names	Freq
192	CT62	426
697	RP11.404K5.2	399
227	CTD.2532D12.4	297
648	RP11.23E10.4	277
326	HAND2	265
816	RP4.694A7.2	264
656	RP11.274B18.4	260
220	CTD.2266L18.1	225
644	RP11.202D1.2	223
923	TMEM207	222

Fig.11| Bootstrapping of 500 iterations to select important genes for stages 1 to 4 (top left, top right, bottom left, bottom right) of the disease

Following the same procedure of 10 fold-cross validation explained in the method section, we chose the model that performed the best among three models - multinomial lasso regression, random forest, and XG boost. Among three models, multinomial regression with lasso penalty had the best performance of 60.2% and 71.1% for predicting for all stages and binary classification respectively. As shown in the Fig.12, in both classification problems, our model predicts stage 1 with high accuracy but predicts stage 4 with low accuracy.

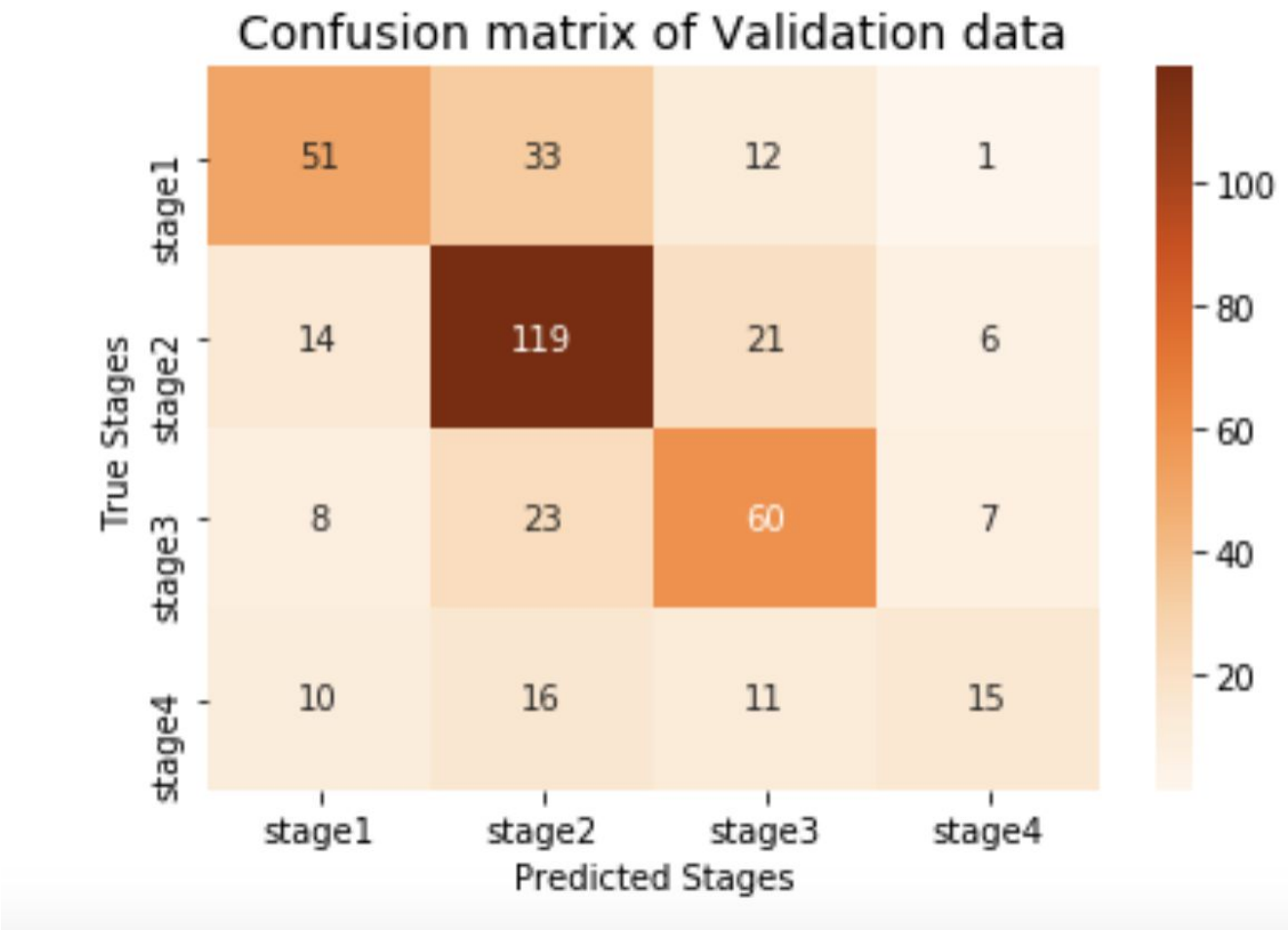

Fig.12|a Confusion matrix for predicting all 4 stages in validation set

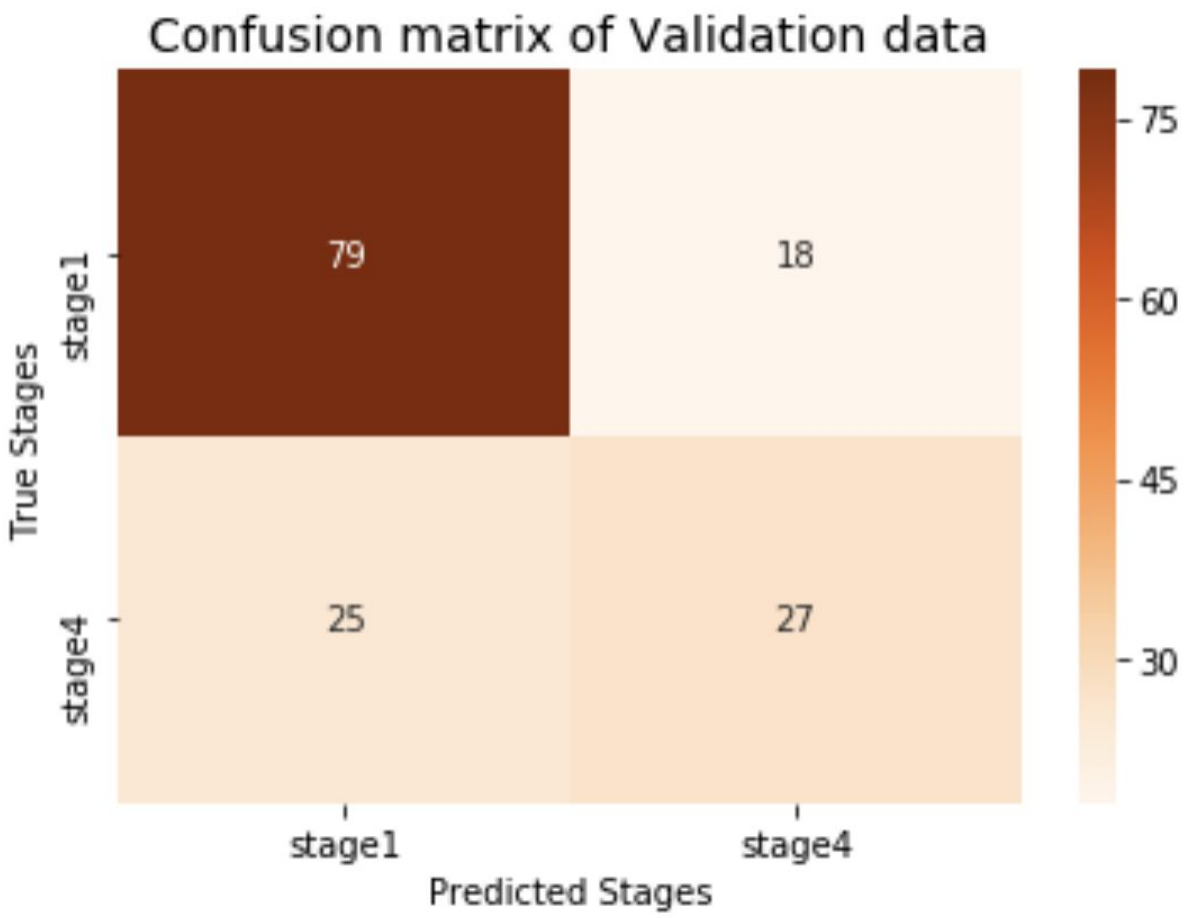

Fig.12|b Confusion matrix for predicting stage 1 vs stage 4 in validation set

Now, we touched hold-out data of 46 patients, and the model predicted with 34.8% accuracy for all stages, and 47% for the binary classification.

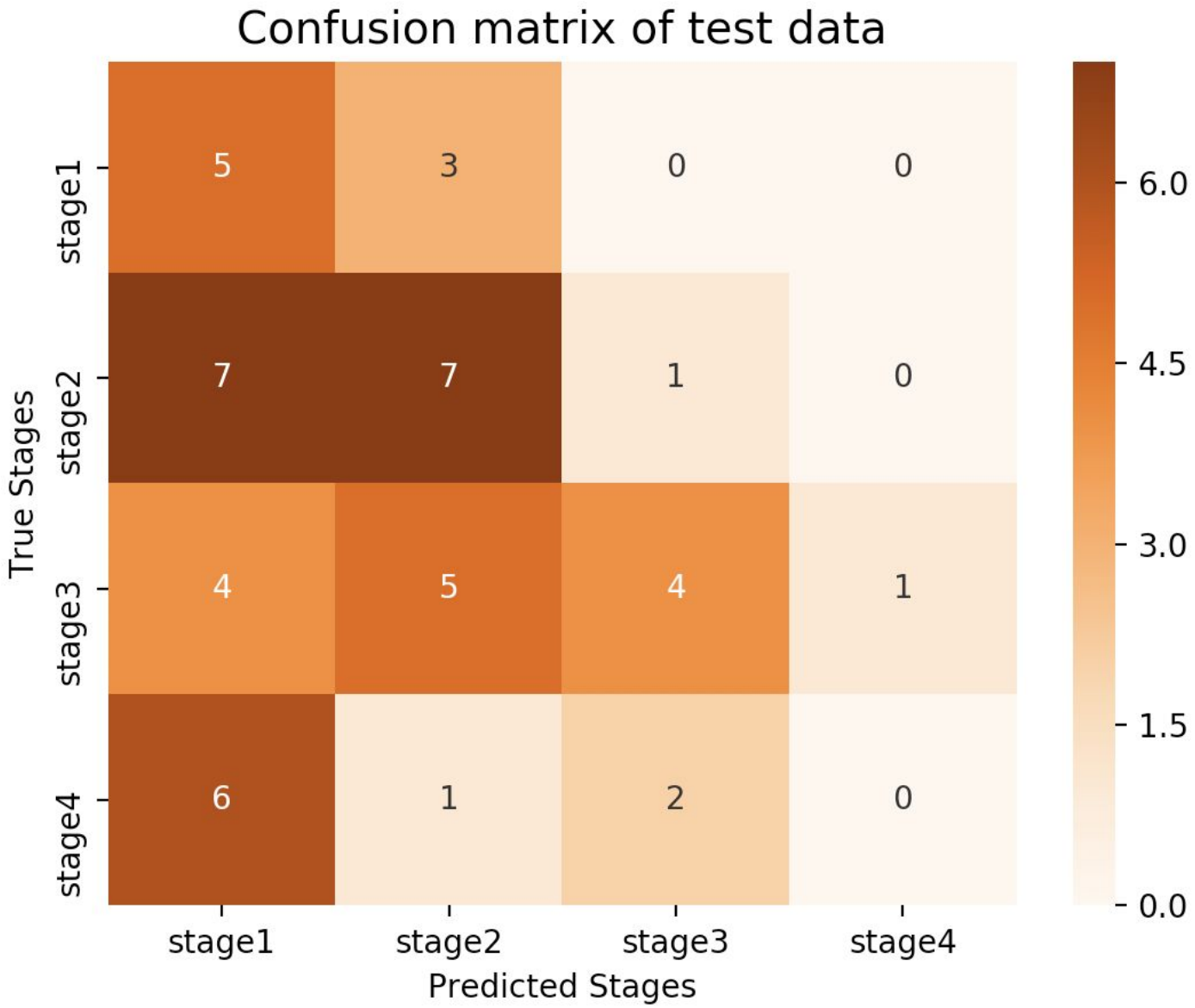

Fig.13|a Confusion matrix for predicting all 4 stages

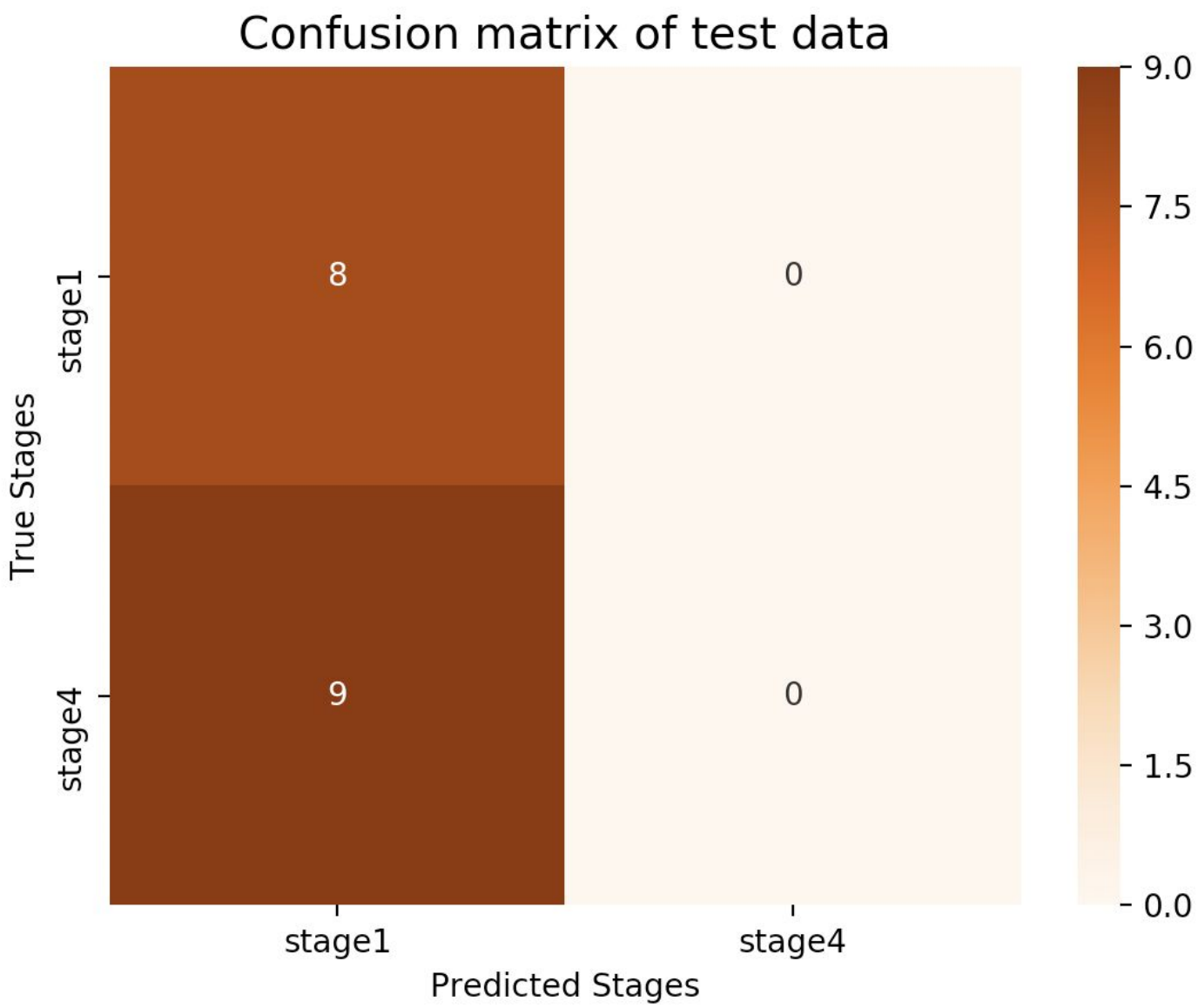

Fig.13|b Confusion matrix for predicting stage 1 vs stage 4 in test set (multinomial regression with lasso penalty)

In both cases, The model could not predict any of the patients in stage 4. One reason is that we have small samples for stage 4, but we analyzed further to understand why the model cannot predict stage 4 with high accuracy. As shown in the Fig.14, 15 most important genes that are selected by the model have low expression levels, and these genes do not have much power to classify patients to be in stage 4.

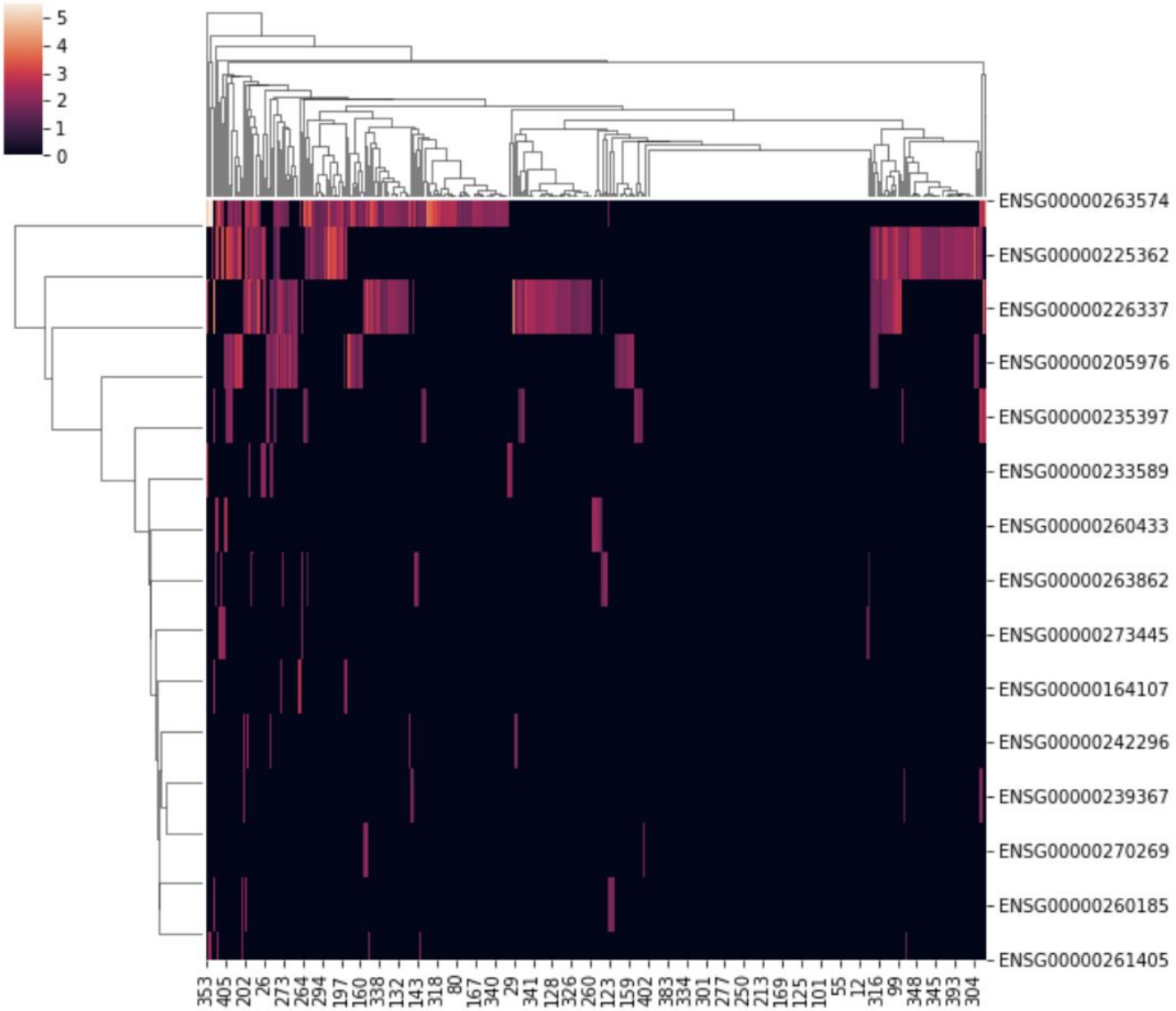

Fig.14| Cluster heatmap of 15 most important genes selected by multinomial lasso regression

Shown in the Fig.14, We also looked at the correlation matrix of 40 important genes in stage 4 to see if genes are correlated to each other. Most genes have correlation coefficient above 0.8. This is another challenge that the model could not choose genes that are independent to each other.

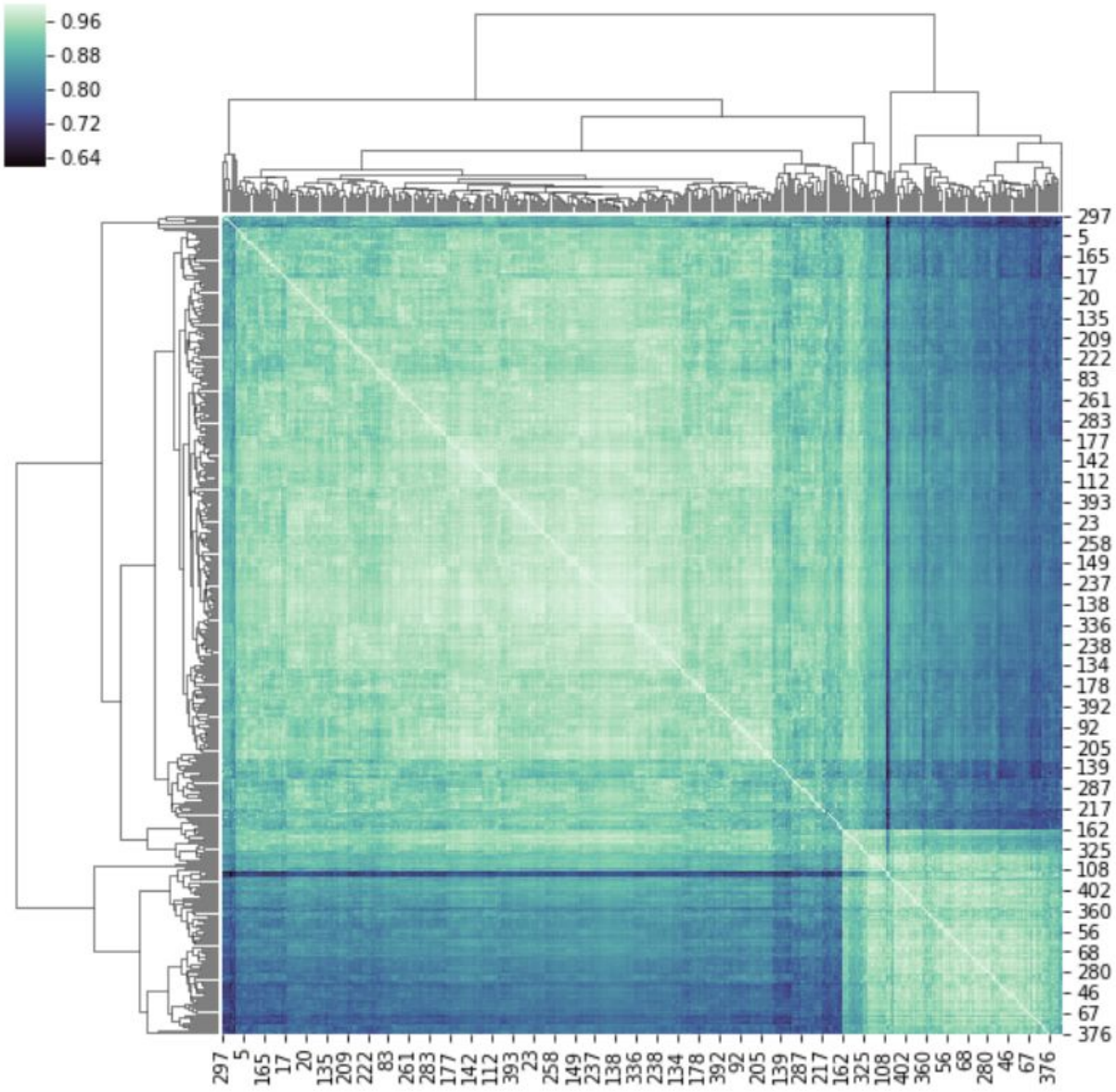

Fig.15| Correlation matrix of 40 most important genes selected by multinomial lasso regression

Pathway analysis:

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. And performing pathway analysis can analyze data obtained from high-throughput technologies, detecting relevant groups of related genes that are altered in case samples in comparison to a control.

Therefore, for our project, after getting genes associated with AMD stages, we can use PAs to find related genomic pathways for each stage. The method we use is GO analysis. The input for Go pathway analysis is 300 genes selected from each disease stage after multinomial lasso model. Unfortunately ,we got exactly the same pathways for each stage(Fig.16)**.** Although some pathways involved in immune response seem to be important, these pathways are not involved in known AMD pathways. The reason for this abnormal result is that the 300 genes we choose have very low overlaps with the GO database. Therefore, after performing GO analysis, GO term has only 1 or 2 genes from the list, which makes the pathways we found look the same.

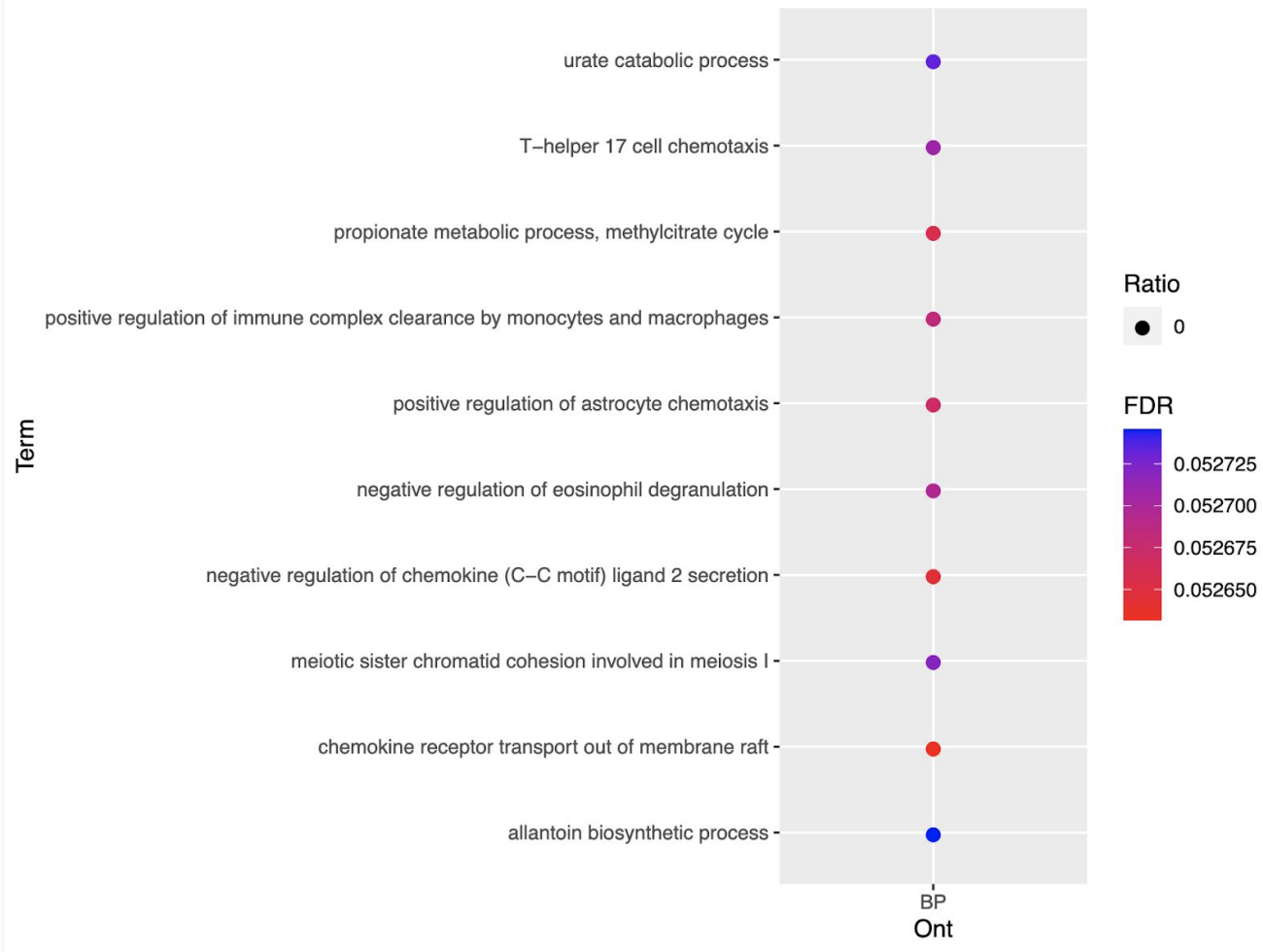

Fig.16| Top 10 pathway after GO analysis

Conclusion

With nearly 58,000 genes, it was challenging to identify important genes that are predictive of different stages. Through low variance filtering and ANOVA analysis, we statistically removed less signaling genes, reducing our feature set drastically to ~18,000.

Furthermore, we found significant genes through various feature selection methods including Multinomial Regression, Random Forest, and XGBboost, achieving the highest training set accuracy of 60.2%. While the stability of our models and accuracy of predictions remained somewhat low (discussed in the next section), we discovered that some genes we found are known to play a significant role in brain damage, leading to eye vision loss. We also looked into genomic pathways to see where in the pathway is disturbed by differentially expressed genes.

Our research aims to provide breakthroughs in prediction, prevention and management of AMD in the elderly through transcriptomics. The identification of key biological pathways and connections will allow us to better develop treatment techniques for the disease. Finally, we may be able to provide medical practitioners with the tools to analyse the onset and rate of advance of AMD.

Challenges and Next Steps

We faced several challenges that impacted our results. In particular, we found that we were severely underpowered -- after taking out our test set, we were only left with the data of 450 patients. This meant that when we tried to train on our small dataset using k cross-fold validation, we perhaps overfitted the data. If we did not use k cross fold validation techniques, we were underpowered. Going forward, we hope to find a larger dataset and check whether their performance improves with an increase in number of datapoints.

We also faced a class imbalance while training our models. Rows from patients of Stage 2 constituted 32% of our overall data. Because of this imbalance, Stage 4 constituted 17% of the data. This imbalance also likely skewed our results, especially in our multinomial lasso regression model where stage 4 was predicted the least accurately. In the future, we hope to use methods such as SMOTE (Synthetic Minority Over-sampling Technique) more rigorously to balance our data.

We hope to continue working on this dataset and dive deeper into exploring how we can improve our existing models based on some avenues that we can explore. We want to conduct a more rigorous class-wise analysis, and run our data science pipeline through different initial datasets. This will help us better understand what may be the cause of our relative inaccuracy in our prediction models. Since our initial modeling results were promising and leave scope for more research, we hope to identify better pathways that can give us greater insight into the causal genes of AMD. Given more time, we also want to conduct rigorous network analysis on the pathways we have identified earlier and the new ones we identify.

References

Age-related macular degeneration - Genetics Home Reference - NIH. (2019). Retrieved from https://ghr.nlm.nih.gov/condition/age-related-macular-degeneration

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, *97*(18), 10101–10106. doi: 10.1073/pnas.97.18.10101

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, *8*(2), e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Ma, S., Song, X., & Huang, J. (2007). Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, *8*(1). doi: 10.1186/1471-2105-8-60

Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C., & Chikina, M. (2019). Pathway-level information extractor (PLIER) for gene expression data. *Nature Methods*, *16*(7), 607–610. doi: 10.1038/s41592-019-0456-1

Morgan, D. J., & DeAngelis, M. M. (2014). Differential Gene Expression in Age-Related Macular Degeneration. *Cold Spring Harbor perspectives in medicine*, *5*(8), a017210. https://doi.org/10.1101/cshperspect.a017210

Nguyen, T.-M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, *20*(1). doi: 10.1186/s13059-019-1790-4

Olsen, T. W., & Feng, X. (2004). The Minnesota Grading System of Eye Bank Eyes for Age-Related Macular Degeneration. *Investigative Opthalmology & Visual Science*, *45*(12), 4484. doi: 10.1167/iovs.04-0342

Ratnapriya, R., Sosina, O. A., Starostik, M. R., Kwicklis, M., Kapphahn, R. J., Fritsche, L. G., … Swaroop, A. (2019). Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature Genetics*, *51*(4), 606–610. doi: 10.1038/s41588-019-0351-9

Raychaudhuri, S., Sutphin, P. D., Chang, J. T., & Altman, R. B. (2001). Basic microarray analysis: grouping and feature reduction. *Trends in Biotechnology*, *19*(5), 189–193. doi: 10.1016/s0167-7799(01)01599-2

Tuo, J., Bojanowski, C. M., & Chan, C. C. (2004). Genetic factors of age-related macular degeneration. *Progress in retinal and eye research*, *23*(2), 229–249. https://doi.org/10.1016/j.preteyeres.2004.02.001

Yao, D., Yang, J., Zhan, X., Zhan, X., & Xie, Z. (2015). A novel random forests-based feature selection method for microarray expression data analysis. *International Journal of Data Mining and Bioinformatics*, *13*(1), 84. doi: 10.1504/ijdmb.2015.070852

Appendix

Dates	Process	Objective 1	Objective 2	Objectve 3
Done	**Question, Data & Background**	**Goals** Understand the scope of the project and the data itself **Steps** 1. Conduct a literature review of Nature, Cell, etc to find out more about the disease itself. Also learn about SVD which has been suggested in conventional literature as possible modeling technique 2. Understand the transcriptomic background behind AMD	1. Understand "biological pathway" through literature and explore existing pathways for AMD if present 2. Explore methods such as Go Enrichment Analysis, KEGG, Reactome used for pathway detection	1. Learn more about existing network analysis tools (InTact, BioGrid) and their use cases to identify which tool can be used for our analysis
Done	**Data Wrangling + Data Exploration**	**Input** Background knowledge + Raw dataset **Goals** Understand our dataset to identify any artifacts that our suprising. For example, does gender bias our data? Are there any genes that cause immense variation and can be visualized instantaneously? **Steps** 1. Clean data for any XY chromosome genes 2. Clean data for any genes that are not expressed adequately at any stage of the disease 3. Normalize data using log2cpm counts 4. Batch correct for age and other non-biological factors as necessary using existing methods suchs as LIMMA, SVA 5. Visualize using clustering, t-sne, PCA and UMAP to visualize genes that cause the most variation **Output** Structured + cleaned data that has been visualized		

Figure 1: Data Science Pipeline

<table>
<tr><td rowspan="8">Done</td><td rowspan="8">Prepare for Modeling</td><td colspan="3">Input</td></tr>
<tr><td colspan="3">Cleaned data with visual variations seen in the data</td></tr>
<tr><td colspan="3">Goals</td></tr>
<tr><td colspan="3">Structure data for single class/multi class classification</td></tr>
<tr><td colspan="3">Steps</td></tr>
<tr><td colspan="3">1. One-hot encode all classes
2. Create combinatorial subsets of the dataset with corresponding with the disease stages</td></tr>
<tr><td colspan="3">Output</td></tr>
<tr><td colspan="3">Primary + secondary tables that can be used for binary and multi class classifiers such as SVM, Random Forest, Lasso Regression</td></tr>
<tr><td rowspan="12">April 4th</td><td rowspan="12">Modeling + Validation(read column-wise)</td><td colspan="3">Input</td></tr>
<tr><td>Objective 1</td><td>Objective 2</td><td>Objective 3</td></tr>
<tr><td>Primary + secondary tables of gene expressions with different classes that can be used for binary and multi class classifiers</td><td>List of genes that are the most differentially expressed among different stages of AMD</td><td>Directional graph of genes connected to each other that together constitute the pathway for different stages of AMD</td></tr>
<tr><td colspan="3">Goals</td></tr>
<tr><td>Construct models to identify sets of genes that can be linked to pathways</td><td>Identify pathways that exist in different stages of AMD and throughout the progression of the whole disease</td><td>Find networks that the newly identified pathways can be fit into and explained through</td></tr>
<tr><td colspan="3">Steps</td></tr>
<tr><td>1. Use binary class classification to find most differently expressed genes between pairs of different disease stages
2. Use multi-class classifiers like Random Forest, SVM to find most differentially expressed genes among all stages of the diesease
3. Track which genes are expressed throughout all stages and compile a dataset of these genes
4. Validate results and go back to previous step if necessary</td><td>1. Perform GO Enrichment Analysis to identify pathways using genes identified in binary-class classification
2. Perform GO Enrichment Analysis for genes identified in multi-class classifications to identify pathways
3. Track to see whether there are common pathways that can be modeled into existing networks</td><td>1. Perform KEGG Network Analysis to explain pathways found from modeling in Objective 2</td></tr>
<tr><td colspan="3">Output</td></tr>
<tr><td>List of genes that are the most differentially expressed among different stages of AMD associated with its importance</td><td>Directional graph of genes connected to each other that together constitute the pathway for different stages of AMD</td><td>Directional graph of pathways that are connected and interact with each other to produce AMD</td></tr>
</table>

Figure 2: Data Science Pipeline (continued)