Genetic Aging Analysis

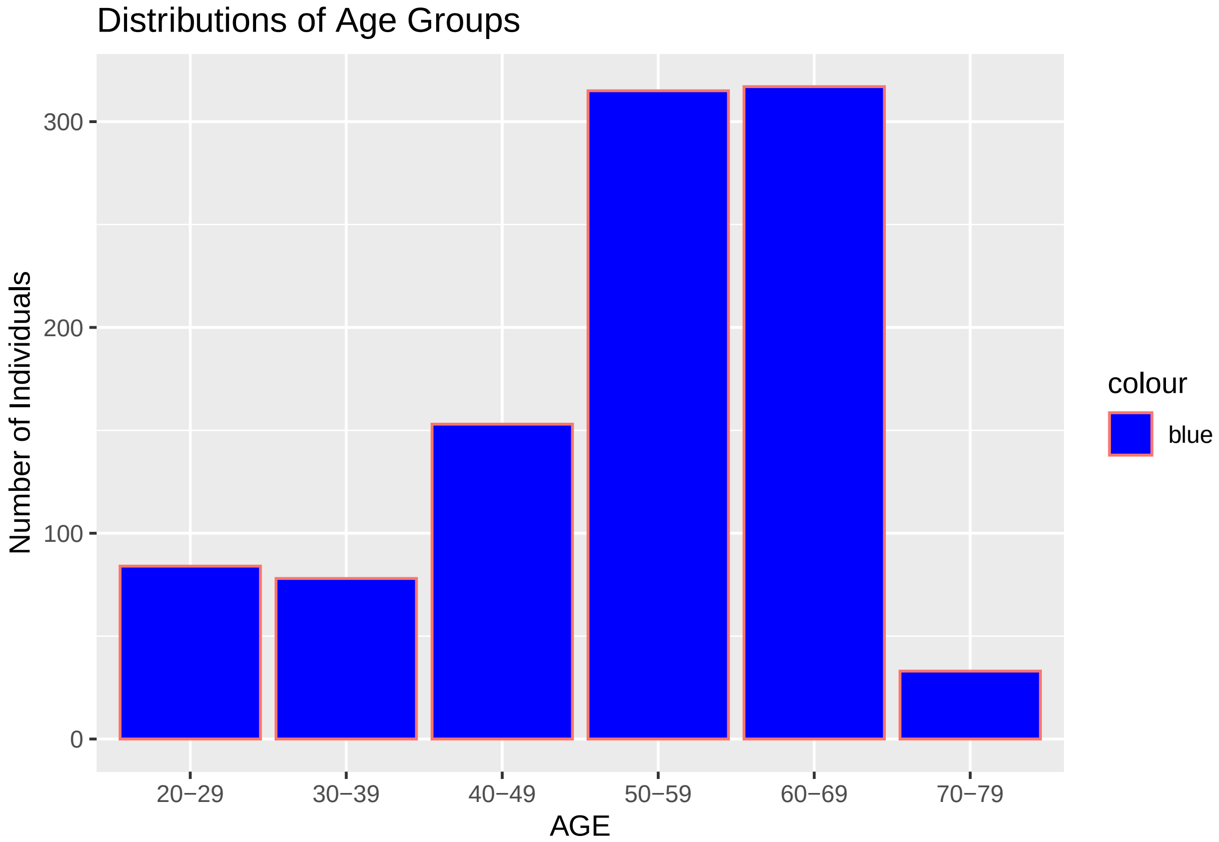
Shane Putney

June 17, 2020

Set Up and Conditions

The goal of this project was to look into the mechanisms of genes on the process of aging. Working in a fashion similar to the workforce, the project was started on June 15th, 2020 and was given a deadline to complete by June 17th, 2020. This creates a time-sensitive environment, resulting in analysis that is more common than that of the 8 month R&D team's analysis.

The data was collected from the gtex portal website, and was processed through R's data.table framework. The genetic data was very wide in size (56,000+ variables) and had nearly 18,000 genetic samples. Out of these genetic samples, there were 980 different individuals assigned to the samples. Each individual had a labeled age range broken up and distributed in the graph below:



The data came as three different files. One listed the gene expression value (as rows) and the sample IDs as columns (56,200 x 17,384). The second data frame came as a file named attributes and contained 22,951 sample IDs and the attributes related to their sampling such as tissue type (22,951 x 63). The third dataset came as a smaller file, listing the first few parts of the sample ID as a column to identify the individual as well as some information about age (920x4).

ETL Process

Coming into the problem agnostically, the goal was to start with some light ETL work between all three data frames. The first issue is the in-memory loading of the gene expression data. To deal with the somewhat larger datasets at hand without having to send it to a database server and extracting it with SQL or dbplyr commands, the data.table library was used. This allowed a much quicker read-in and manipulation of the data from within the R environment when using the libraries available tools (f the dataset had been larger or more time was allotted on the project, setting up an Amazon RDS MySQL database and querying from there would have been an option as well).

The second issue was gathering the data into a format where the data would be expressed as a matrix of [AGE|Gene Expression]. The two issues regarding this was the un-matched sample ID's when joining and each individual having multiple tissue samples of gene expression. The first issue can be shown below:

```
##   Gene_Expression_SampleID Phenotype_SampleID
## 1 GTEX-1117F-0226-SM-5GZZ7         GTEX-1117F
## 2 GTEX-1117F-0426-SM-5EGHI         GTEX-1117F
```

To use an inner join combining on these columns a form of RegEx must be used on the Gene Expression sample ID's before joining them. Using this implementation alone though is a poor method of set-up, as the samples should be independently sampled. To avoid this further down, each sample is also broken down by tissue type. For example, a new dataset may be named Muscle_Tissue with the desired format of [AGE|Gene Expression] and contains only samples that were taken from muscle tissue. This ETL process was simplified into functions, which can be found in the github's "functions" folder. The following tissue types were taken into consideration.

Tissue Types:

- Adipose Tissue
- Blood
- Blood Vessel
- Brain
- Colon
- Esophagus
- Heart
- Lung
- Muscle
- Nerve
- Skin
- Thyroid

Creating Predictive Models for Inference

The next step was to create seperate predictive models for each tissue type that would predict the age category. In an ideal world the implementation would be that of a regression, but due to the nature in which the data was collected the best methodology would be classification. Misclassifications are okay for this model, as a misclassification could indicate a person aging faster/slower than average within that tissue type. As a hypothetical, there exists a 52 year old who runs every day. The algorithm may say the heart tissue in this man belongs in the 30-39 category, but other tissues below in the 50-59. This would be an IDEAL misclassification, where the model did a great job of generalizing.

The original method of classification proposed was going to be a basic feed forward neural network implemented in Keras, and then using the iml package in R to grab important variables. Unfortunately, even after changing different hyper-parameters such as learning rate, loss, class weights, drop layers, number of layers, batch size, epoch (number of cycles), or neurons per layer the model always seemed to grab a local minimum based on initial weights and predicted a single class 100% of the time.

The second proposed method of classification was to set up the age groups as a regression problem, and use lasso for variable selection. The next step is to convert the problem back to a classification problem and tune/train an XGboost model (on the selected variables) using the tidymodels framework. This method was much more direct, and gave results good enough for analysis. The file describing the method may be found in the "functions" folder under the file name "genetic_tissue_age_model.R". All scripts that combine the ETL and modeling process can be found in the "Models" folder. The following approximated classification rate were found for each tissue type.

Tissue Models: Validation Classification Rate

- Adipose Tissue: 0.83

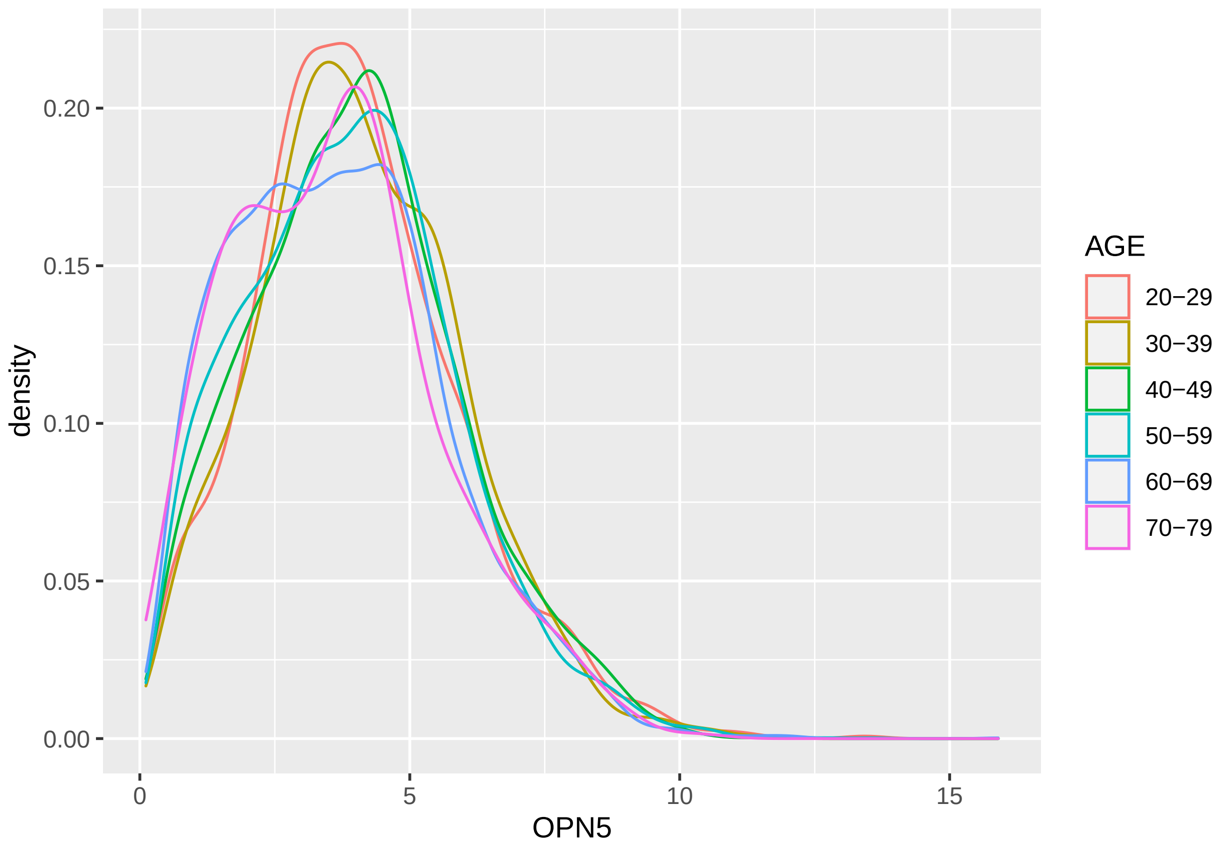
- Blood : 0.89
- Blood Vessel : 0.81
- Brain : 0.68
- Colon : 0.92
- Esophagus : 0.78
- Heart : 0.88
- Lung : 0.93
- Muscle : 0.90
- Nerve : 0.95
- Skin : 0.73
- Thyroid : 0.92

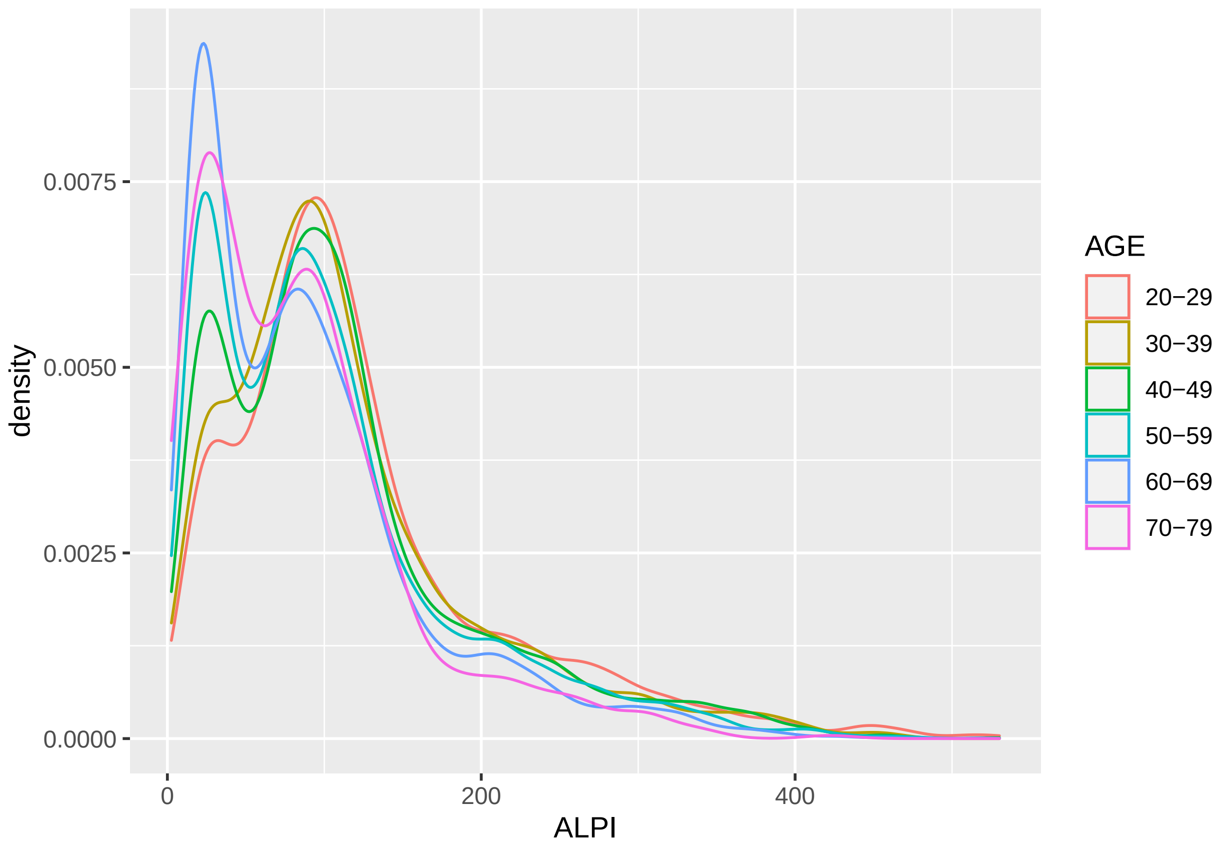
Analysis

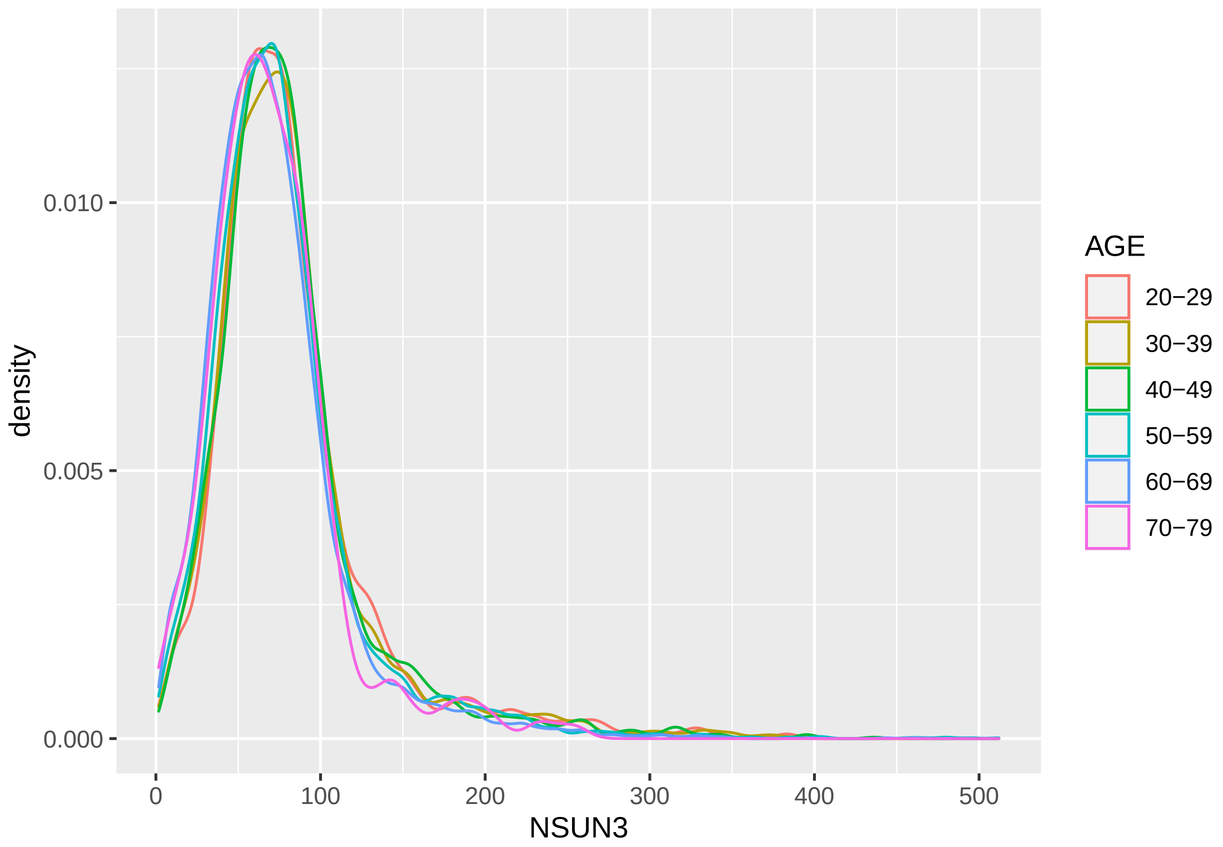
The goal of this analysis is to take a look at the individual genes that have the most impact throughout all the models. The following table shows a list of genes and the number of times they have were selected among the 12 different tissue models (grabbing the names of the genes which have at least 5 intersections).

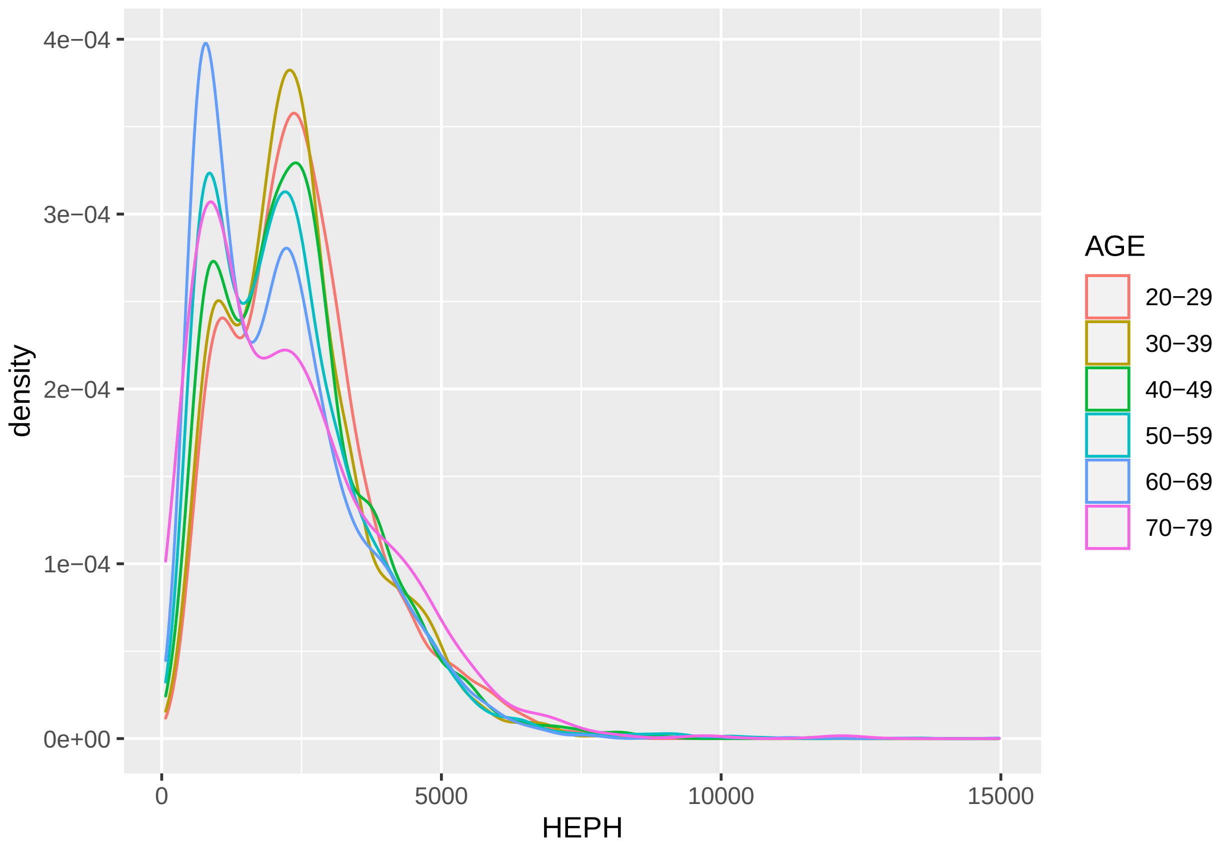
```
##                            WDR1 GABRP OPN5 ALPI NSUN3 PCDHB1 RPS12 MTAP ZNF384
## Number of Intersections    8     8    8    7     6      6     6    6      6
##                            HEPH VWA3B HSD17B4 PCDHB8 USP17L8 CALHM2 MGST1 ZNF257
## Number of Intersections    6     5       5      5       5      5     5      5
##                            KLK1 NLRP7 CH507.42P11.8 RFPL1 FAM9A SPACA5B SH3BGRL
## Number of Intersections    5     5             5     5     5       5       5
```

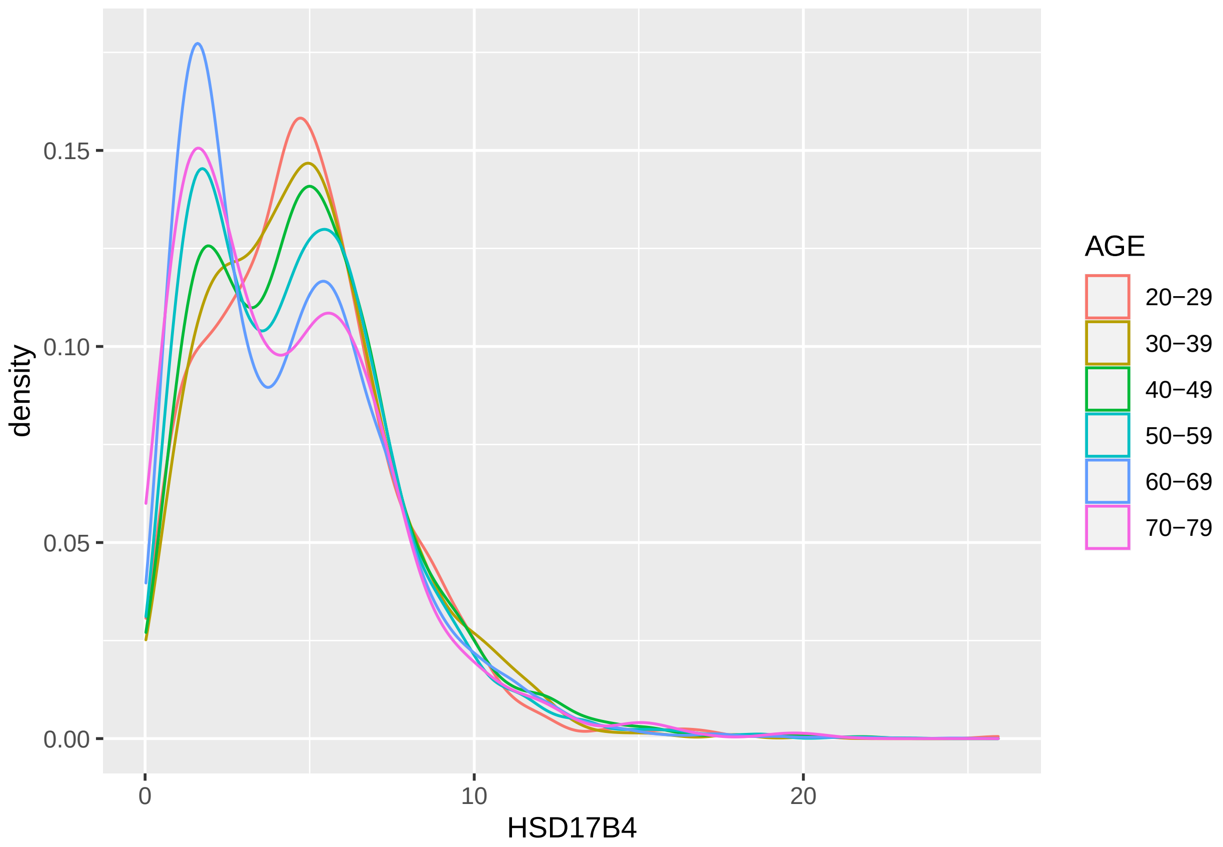
One thing to note is that not a single gene was selected for all 12 models. Albeit discouraging, the next step is to take a look at the distribution of some of the most prominent genes. The genes that look the most promising are showcased below with their respective density plots (seperated by Age).

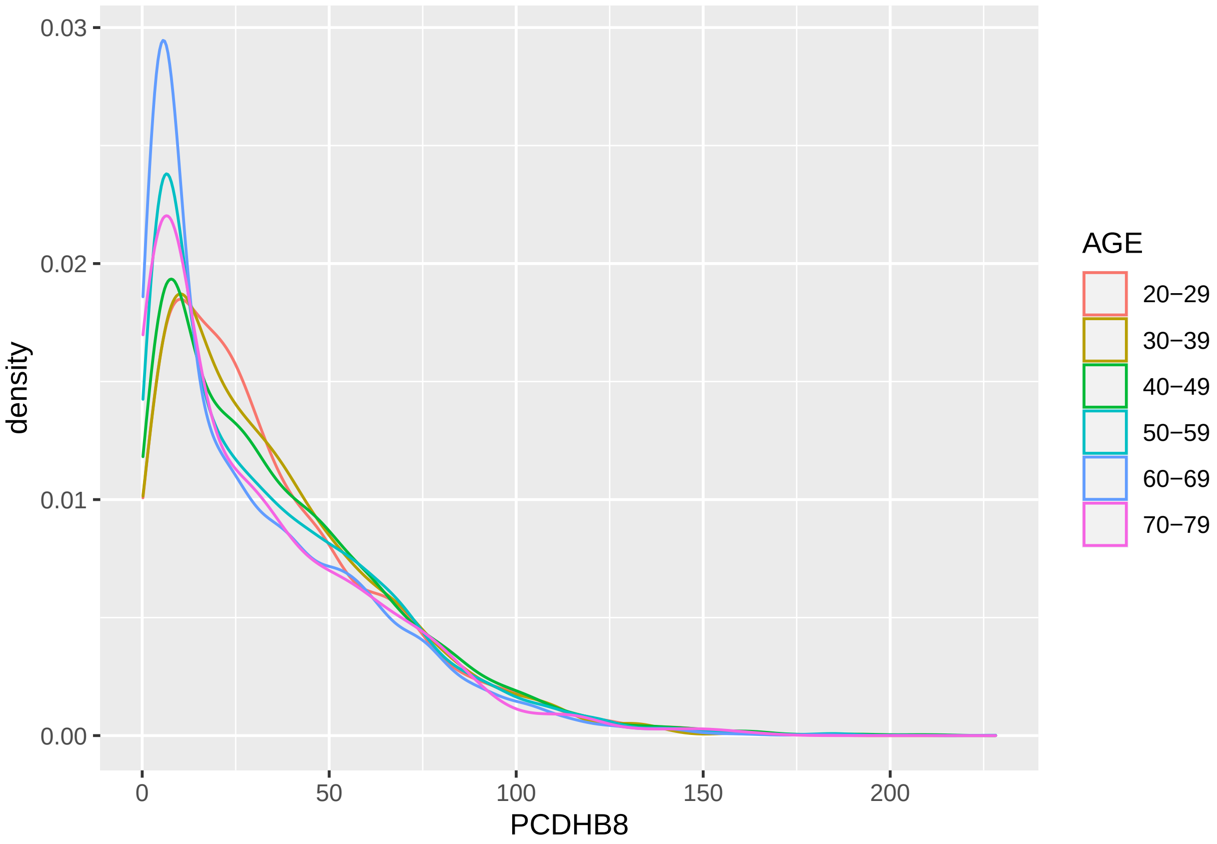

density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
OPN5
AGE
20−29
30−39
40−49
50−59
60−69
70−79

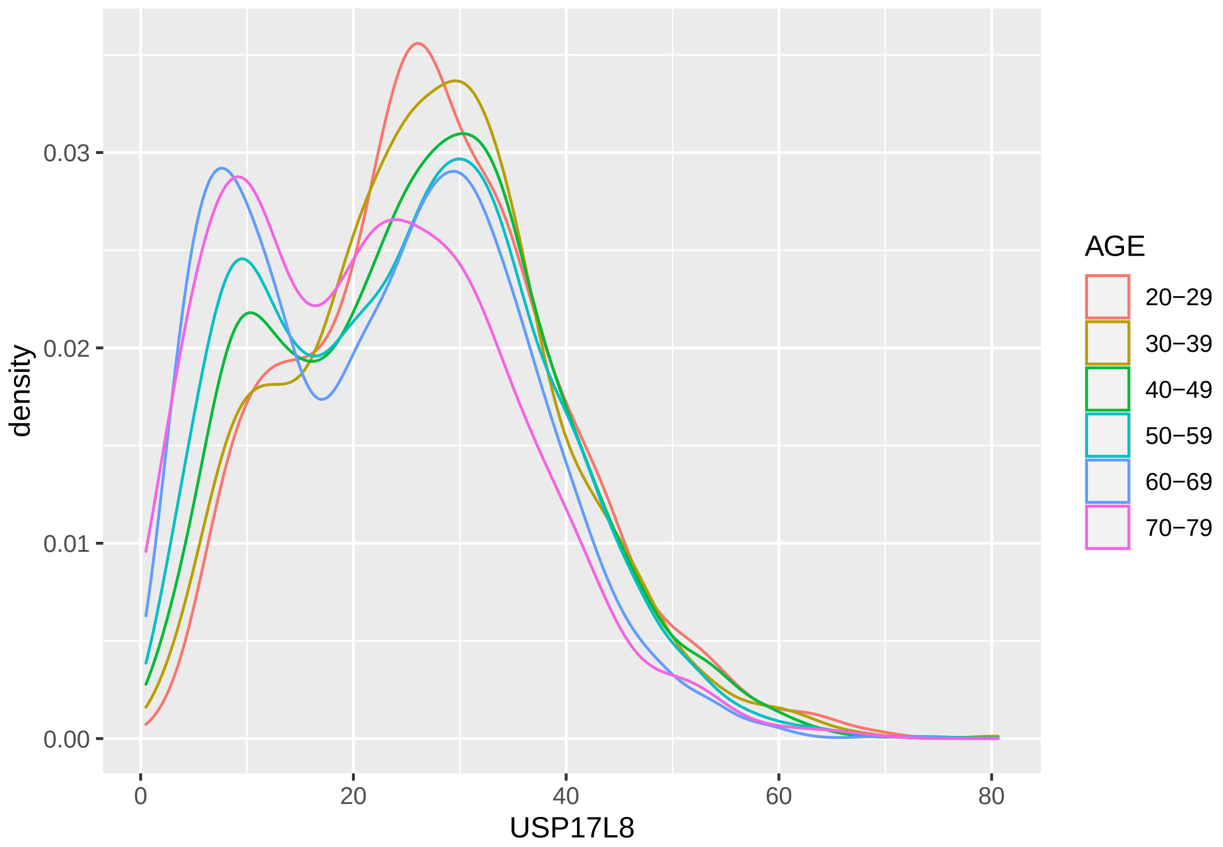

density
0.0000
0.0025
0.0050
0.0075
0
200
400
ALPI
AGE
20−29
30−39
40−49
50−59
60−69
70−79


density
0.000
0.005
0.010
0
100
200
300
400
500
NSUN3
AGE
20−29
30−39
40−49
50−59
60−69
70−79


density
4e−04
3e−04
2e−04
1e−04
0e+00
0
5000
10000
15000
HEPH
AGE
20−29
30−39
40−49
50−59
60−69
70−79


0.15
0.10
0.05
0.00
density
0
10
20
HSD17B4
AGE
20−29
30−39
40−49
50−59
60−69
70−79


0.03
0.02
0.01
0.00
density
0
50
100
150
200
PCDHB8
AGE
20−29
30−39
40−49
50−59
60−69
70−79



Conclusion

With some quick research through the gene cards website, the common pathways that the variables belong to are:

Pathways:

- Metabolism of Proteins
- Deubiquitinating
- Mineral Absorption
- rRNA processing in the Mitochondria
- Synthesis of Bile Acids and Bile Salts
- Circadian Rhythm Related Genes

The next step of the research would look towards a biological mechanism for the genetic differences in aging. Understanding what is happenening on a biological level would be a great step in the right direction, and hopefully this paper helps give an understanding of where to start.