Performing Gene-Gene Correlation Analysis Across Three Human Age Groups to Improve Biological Age Prediction Models

Tycho Grapendaal[1]
Responsible Professor: Marcel Reinders[1]
Supervisors: Bram Pronk[1], **Inez den Hond**[1], **Gerard Bouland**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Tycho Grapendaal

Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Bouland, Kaitai Liang

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Aging is the biological process that changes the body over time. When we age our bodies become more prone to disease and other health risks. But not everyone experiences these changes at the same age. This is because the age of our cells (biological age) does not always match our chronological age (time since birth). Being able to predict someone's biological age and comparing it to their chronological age can be used to infer if someone is indeed more prone to diseases or other health risks.

Other studies have been able to predict the age of cells by using gene expressions. They explore the number of expressions in young and old individuals to identify genes that are affected by age. What has not yet been explored is how the correlation of gene pairs are affected by age. How genes cooperate can change with age, this can be captured by looking at how genes correlate and how that correlation changes with age. This paper will explore these correlations and answer the following question. By performing a correlation analysis between features of young individuals, and on the same features for old individuals, can we interpret any differences and use those to improve current age prediction models?

During this study we found a lot of gene pairs that have a significant difference in correlation from younger to older individuals. We also identified hub genes that change correlation with many other genes. Using these genes to train a linear regression model we were able to predict the age of cells with a Mean Absolute Error of 9.7835.

Using the hub genes we were not able to improve the current existing linear regression model. But we did identify genes that have earlier been linked to aging. Like LIMD2, but also a lot of ribosomal genes and mitochondrial genes, both of which lose functionality with aging.

1 Introduction

The human body endures significant mechanical, chemical, and environmental stress on a daily basis. Minor injuries such as cuts, bruises, and muscle strains are very common. The same goes for diseases that make us sick and weaken our immune systems when they fight back. Yet, in youth our cells repair very quickly and we heal completely from this damage. However, as we age, our cells become less active, stem cells do not work as well, and our body's natural repair systems slow down, making it harder for tissues to stay healthy. A once minor injury can now result in prolonged inflammation, delayed wound closure, or chronic pain. The immune system also weakens with age. A normal flu can be deadly in old age as thousands of people die from it every year, most of whom are 65 years or older [4]. But this does not happen to everyone at the same age. There are many 65-year-olds in great shape, so age does not say everything. This is because there are two types of ages, the chronological age, which is the time since birth, and the biological age, the age of our cells. The biological age is responsible for the effects mentioned above and knowing someone's biological age can tell us if someone is healthy or not, whether they are more prone to diseases, and if they need more medical care.

One way to research biological age is to look at the expression of genes in young and old individuals. Certain genes have a negative correlation between their expression and the age of the individuals. This means that the older our cells get the less these genes are expressed. There is already a lot of research done on biological age, and there exist models that can predict someone's biological age based on their gene expressions [14]. Previous studies have also already explored genes that are more affected by aging. For example, Soheili-Nezhad et al. showed how gene length negatively correlates with their expression in

older individuals [9]. Additionally, Olga Ibañez-Solé explored age-related changes in gene expression, including correlations between average expression levels in young and old cells [5]. They also identify a correlation between the length of a gene and their expression over age.

These studies looked at gene expression for single genes, and explored how these expressions change with age. But they did not look at gene pairs and how they work together. Do correlations between genes change with age and are these changes significant? Are there genes that act as hubs and change correlation with many other genes across age groups? By performing a correlation analysis between features of young individuals, and on the same features for old individuals, can we interpret any differences and use those to improve current age prediction models?

To fill this gap and answer these questions, we calculated the correlations between genes of different types of blood cells. This has been done for three different age groups. Then we looked at the differences between these correlations to identify gene pairs that shift from a strong to a weak correlation (or the other way around) from one age group to another. We also created networks of those genes to look for hub genes that change correlation with many other genes. We then used these hub genes to train an age prediction model.

2 Methodology

2.1 Data

The dataset consists of single-cell RNA sequencing measurements profiling gene expression in 1,058,909 blood cells, covering 36,169 genes. For each cell, expression levels are quantified by counting unique mRNA transcripts per gene. It is a version of the Asian Immune Diversity Atlas (AIDA). There are 508 healthy donors from Asian countries like Japan, South Korea and Singapore between the ages of 17 and 75. The data is hosted by The Chan Zuckerberg Initiative and can be downloaded using this link https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508. Select the first dataset "AIDA Phase 1 Data Freeze v1: Chinese, Indian, Japanese, Korean, and Malay donors in Japan, Singapore, and South Korea". This dataset was also used by Enikő Zakar-Polyák et al. [14]. They use the data to train single cell transcriptomic clocks using ElasticNet regression models. We used the same data to find correlations between genes and train the same models on genes with many significant differences in correlation across age groups.

2.2 Age Ranges

During this research the correlation coefficients are calculated for three age groups: young, middle-aged and old. To define the age ranges for the three age groups we looked for age ranges that closely matched biological age groups like early adulthood (20-30), middle adulthood (30-65), old age (65+) [13], while still keeping at least 50 donors (for most cell types) to ensure that all correlations have a small p-value. This resulted in the following age ranges, young (19-30), middle-aged (40-50) and old (60-75). The age groups have 10-year gaps between them. We chose to do this to make it easier to see any differences (if there are any) between the correlation coefficients between these groups. The gaps make sure that the groups are not that similar.

The age ranges are based on the total distribution of the donors' ages for the entire dataset. The age ranges are also the same for all cell types to make comparisons between

the cell types easier. Figure 7 shows the distribution of the donors' ages for the entire dataset. In the final age groups, there are 94 young, 148 middle-aged, and 55 old donors.

2.3 Cell Types

The dataset used during this research has a total of 33 different cell types. When making subsets of the data, we chose 9 cell types instead of all 33. We chose to do this because we wanted to reduce the total number of results, so it would still be feasible to analyze them within the time of the project. To still get the best results possible, we chose the 9 best performing cell types based on the prediction-models that were created by Enikő Zakar-Polyák et al. [14]. These cell types are CD8-positive alpha-beta T cell, CD8-positive alpha-beta memory T cell, CD4-positive alpha-beta T cell, central memory CD4-positive alpha-beta T cell, effector memory CD4-positive alpha-beta T cell, gamma-delta T cell, regulatory T cell, double negative T regulatory cell, and innate lymphoid cell.

2.4 Preprocessing

The preprocessing consists of creating subsets for every cell type. First, all the cells that are of a different cell type are removed. Then the genes that have a very small expression count are removed. A small expression count is considered less total expressions than the number of cells. So, if there are 3000 cells in the subset, all genes with less than 3000 transcripts in total are removed. For the next step, the average normalized expression (the log-normalized expression value $\tilde{x}_{ij}$ for gene j in cell i was calculated as $\tilde{x}_{ij} = \ln\left(1 + \frac{c_{ij}}{\sum_k c_{ik}} \times 10^4\right)$, where c_{ij} is the raw count of gene j in cell i, $\sum_k c_{ik}$ represents the total counts in cell i, and ln denotes the natural logarithm) of the genes is taken for all the cells of the same donor. So, the matrix goes from cells by genes to donors by genes. To remove unwanted outliers, donors who have less than 10 percent of the median number of cells are removed. So, if the median number of cells for the donors is 20, all donors with 2 cells or less are removed. The final step of the preprocessing is to make subsets for the age ranges. There are three subsets created: young (19-30), middle-aged (40-50) and old (60-75).

When all the subsets were created we looked at the distribution of the donors' ages for all the subsets and the average number of cells for all of the ages (Figures 1 and 8). Based on this information, we excluded CD8 positive alpha-beta T cells, double negative regulatory T cells, and innate lymphoid cells from the results since they did not have enough data for the results to be reliable. This results in six cell types for the rest of the results.

2.5 Correlation Analysis Across Age Groups

The Pearson's correlation coefficients are computed for all gene pairs in the dataset after it has been preprocessed. This was done for the six cell types mentioned above and in the three different age groups. This results in 3 different correlation coefficients for every gene pair, one for each age group.

Then the log fold change is calculated between young-middle-aged, middle-aged-old and young-old. $\text{LFC\_young\_middle} = \log_2\left(\frac{\text{young\_correlation}}{\text{middle\_correlation}}\right)$. The difference between the correlation is also calculated and the Fisher z-transformation is used to check if the difference is significant.

The Fisher z-transformation converts a Pearson correlation coefficient r to an approximately normally distributed variable z:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \tag{1}$$

To test whether two correlation coefficients r_1 and r_2 are significantly different both normally distributed variables are calculated.

$$z_1 = \frac{1}{2} \ln \left(\frac{1+r_1}{1-r_1} \right) \tag{2}$$

$$z_2 = \frac{1}{2} \ln \left(\frac{1+r_2}{1-r_2} \right) \tag{3}$$

Then the difference between the variables and the standard error of the difference are calculated. Where n_1 and n_2 are the sample sizes for each correlation.

$$z_{\text{diff}} = z_1 - z_2 \tag{4}$$

$$SE = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{5}$$

Then for the final step the difference is divided by the standard error of the difference to get the final normally distributed value which can be turned into a p-value.

$$z = \frac{z_{\text{diff}}}{SE} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{6}$$

We consider a difference between correlation coefficients to be significant if $|z| > 2.576$, because that corresponds to a p-value of 0.01.

Gene pairs were identified based on the following criteria: (1) a significance threshold of $p < 0.01$, (2) an absolute $\log_2$ fold change ($|\text{LFC}| \geq 0.7$) across all correlation differences, and (3) a strong correlation ($r \geq 0.8$) in at least one age group. These strict thresholds were selected to detect pairs that shift from a strong correlation ($r \geq 0.8$) to weak correlation ($r \leq 0.1$), or the other way around.

2.6 Creating Networks of Genes

A network of genes for each cell type has been constructed to see how many correlations change for a single gene. For example, some genes might lose correlation with only a single other gene from young to old, while others might lose correlation with 100 or more. We create a network (for every cell type) where all the genes become nodes. Two nodes are connected with an edge if the difference between the correlation coefficients is significant between the young and the old group, and the absolute log fold change between the young correlation and old correlation is at least 2. These thresholds were chosen to look for genes that have significant differences with other genes that are also an increase (or decrease) of 4 or higher. This means that nodes are connected if they go from a weak (0.1-0.2) to a strong correlation (0.6-0.9) or the other way around.

The hub genes are then identified by looking at the node degree distributions and fitting a power law to validate the results [6][8]. A gene is considered a hub if they are connected to at least 10 percent of the other genes. The hub genes across cell types are identified as

well. This is done by adding the degrees (connections) of the genes together for all the cell types, except for the central memory CD4-positive, alpha-beta T cell type. This cell type is excluded because it does not follow the power law. The threshold for hubs across the 5 cell types is set at 300 connections. 300 was chosen since the average number of nodes in the networks is 600, (600 * 5)/10 = 300.

2.7 Prediction Model

We used the hub nodes found in the effector memory CD4-positive, alpha-beta T cell type to train a ElasticNet regression model. This has been done in the same way that Enikő Zakar-Polyák et al. trained their ElasticNet regression model [14]. We applied a 5-fold cross-validation using the log-normalized gene expression count in the data. Then there are 5 prediction models that are applied on all the cells of the same cell type and the Mean Absolute Error and correlation between age and predicted age are calculated to rate the models.

2.8 Code and Environment

All the results shown in this paper can be reproduced by running the python code on this Github page: https://github.com/TychoGrapendaal/CSE3000. All of the results are created by a Jupyter notebook or a python file using Python 3.11.4, Anndata 0.11.4, Numpy 1.24.3, Pandas 1.5.3, Scipy 1.10.1, Scanpy 1.11.1, Matplotlib 3.7.1, Seaborn 0.13.2, Sklearn 1.3.0, Powerlaw 1.5.

All the figures in the Results (see subsection 4) can be replicated by running the files in the experiments folder in the Github repository. There is also a README file that explains step by step how to replicate the results.

One of the code files (apply_clocks_general.py) is made by Enikő Zakar-Polyák et al. [14]. We use their code file to apply the prediction models in the exact same so we can compare performance. We did write the file prediction.py (since this code is not provided by Enikő Zakar-Polyák et al.) but it trains the models in the same way described by Enikő Zakar-Polyák et al. [14].

3 Responsible Research

3.1 Human data

The data used during this project are blood cells from human donors from Asian countries. This is a processed version of the Asian Immune Diversity Atlas (AIDA). This data is public and hosted by the Chan Zuckerberg Initiative. The dataset is anonymized, meaning all personally identifiable information has been removed, and all the donors gave permission. This is crucial to protect donor confidentiality.

The data is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Which means that we are free to use, share and adapt the data under the following conditions. "Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. No additional restrictions - You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits." (Page 1) [2]. Since we published all our results on

Github and allow everyone to use the data under the same conditions and provide credit and links to the original source we follow both of these conditions.

Since all the donors are from Asian countries the results can be biased towards these population groups. We point this out to ensure that we do not draw too general conclusions. Because there are certainly differences between aging when comparing different population groups. For example the life expectancy of Japan is 84 years while in the Netherlands it is 82.

3.2 Predicting Age

Being able to accurately predict someone's biological age can have many ethical implications. Some positives would be that we can help people live healthier lives with preventive care if we can predict their biological age. Knowing more features that influence how we age could help us change our lifestyles for the better. We might also be able to use this information to develop better treatments.

However, there are also possible negative consequences that could come with predicting someone's age. For example, if we know what someone's biological age is and it is much higher than their chronological age, it might be used as a reason to increase their insurance premiums. The models used to predict someone's age might also have a lot of biases because of a lack of data from certain groups. This might cause certain treatments to only be effective on the groups that the models were trained on.

3.3 Reproducibility and Integrity

All the results shown in this paper can be reproduced using the code in the Github repository below https://github.com/TychoGrapendaal/CSE3000. There is also a test dataset that can be used to confirm the correctness of the programs. The data used to create the results can be found and downloaded using this link https://cellxgene.cziscience.com/collections/ced320a1-29f3-47c1-a735-513c7084d508. Select the first dataset "AIDA Phase 1 Data Freeze v1: Chinese, Indian, Japanese, Korean, and Malay donors in Japan, Singapore, and South Korea" (for more information see subsection 2.8).

All results shown in this paper are complete, no information was excluded to force conclusions. All the results can be interpreted by the reader and reproduced if necessary. There are possible biases in the results like the limited races and cell types in the dataset. These biases and limitations are discussed in the Discussion (see subsection 5).

4 Results

4.1 Distribution

The scRNA-seq dataset used for the results contains 1,058,909 blood cells from 508 healthy human donors. They are all between 19 and 75 years old and from the countries South_Korea, Singapore and Japan. The entire dataset has a distribution for the ages of the donors as shown in Figure 7. The age ranges used for the results are based on this distribution. However since there are 33 different cell types in the entire database the distribution might change across cell types. We chose to look at the 9 cell types that resulted in the best-performing cell-type-specific single-cell clocks made by Enikő Zakar-Polyák et al. [14], and Figure 1 shows the number of donors for every age for those 9 cell types. For most cell types,

the distribution resembles the overall dataset (Figure 7). There is only one cell type that differs a lot from the others. This is the lymphoid cell type. The reason that the distribution is so different is because the dataset had only lymphoid cells from around 160 donors instead of around 500 for most other cell types. This means that there are immediately a lot less donors to look at during the correlation analysis which is why we exclude it from the results.

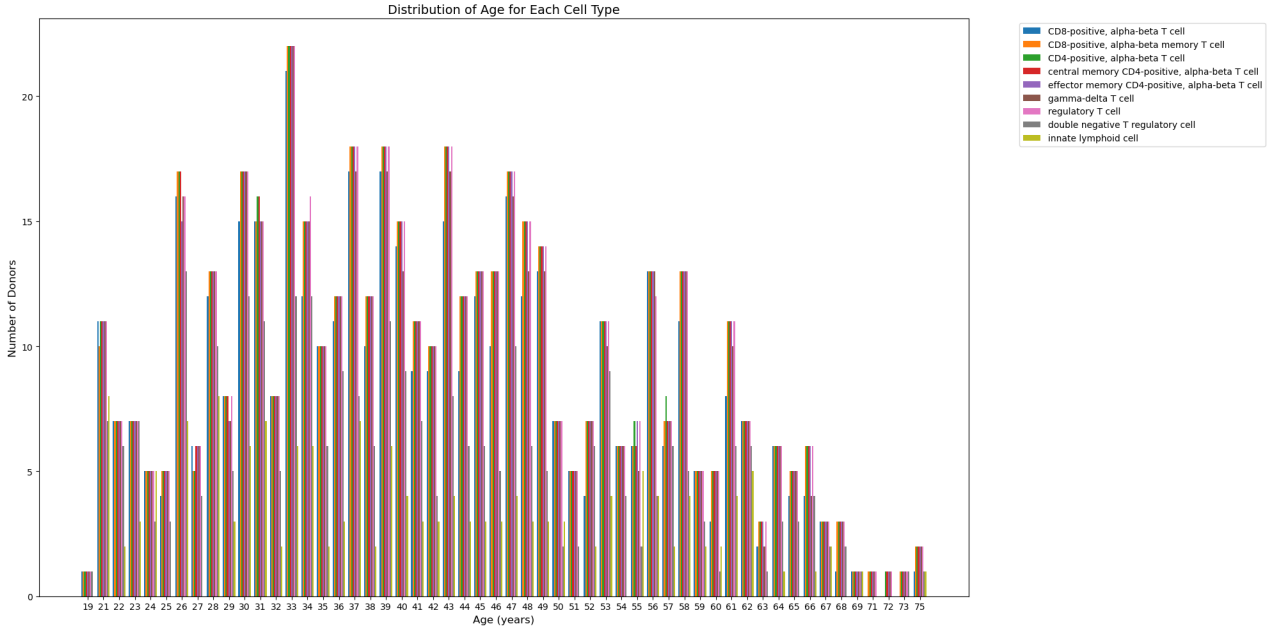

Figure 1: The distribution of the ages for the 508 donors in the dataset for the 9 different cell types. The x-axis shows the age in years of the donors and the y-axis shows the number of donors in the dataset that have this age. There are 9 differently colored bars that represent the different cell types. The legend in the top right shows which color is which cell type.

We also looked at the average number of cells for every age. Figure 8 shows the average number of cells for every age for a specific cell type. The lymphoid, double negative and CD8-positive alpha-beta cell types all have a very small amount of cells for certain ages (multiple ages only have one cell type on average) even after the preprocessing. This means that the results are also not as reliable and these cell types are excluded as well.

4.2 Gene Pairs

To find gene pairs that have different correlation coefficients for every age group that are also statistically significant we calculated the difference between the correlation coefficients from young to middle, middle to old and young to old age groups. We then used the Fisher z-transformation to calculate z-scores for all of the differences and used volcano plots to visualize them (Figure 2). The plots show if a gene pair has significant differences between their correlation coefficients in the three age groups and what their minimum log fold change is between the groups. Every plot shows a lot of points below the red line which means that

they are not significant but there are also points with a very high log fold change that are still insignificant. This is because they have very small correlation coefficients, the young group could be 0.0001 the middle 0.001 and the old 0.01. The log fold change would be very high since the increase/decrease between the correlations is 10, but the absolute difference is still very small and not significant.

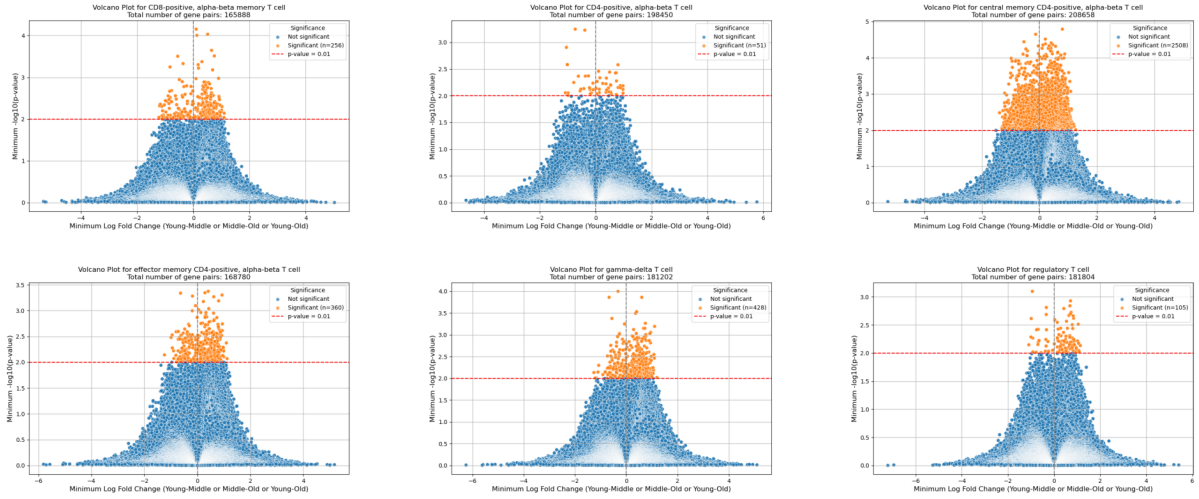

Figure 2: Volcano plots for 6 different cell types. Every point represents a gene pair. The x-axis shows the minimum log fold change between young-middle, middle-old and young-old correlation coefficients. The y-axis shows the minimum -log10(p-value) for the three correlation differences. The reason it shows the minimum is to ensure that all the correlation differences are significant if the gene-pair is above the red line, and that they all have at least the log fold change shown in the plot.

When filtering for gene pairs with a correlation of at least 0.8 in one of the age groups, an absolute log fold change of 0.7 and a p-value of 0.01, we find the genes in Table 1. One of these genes is LIMD2 which has been identified as a cell aging-immune/inflammation-related hub gene by Ping Tao et al.[12]. It starts with a very strong correlation with PFN1 in young individuals and ends with a very weak correlation in old individuals. The high correlation (young age) suggests that LIMD2 and PFN1 are co-regulated or functionally linked in younger individuals. PFN1 (Profilin-1) is an actin-binding protein critical for cytoskeletal dynamics, cell motility, and immune cell function (e.g., T-cell activation) [7]. LIMD2 (LIM Domain-Containing Protein 2) is implicated in cell adhesion, migration, and immune signaling [3]. Their tight correlation in youth may reflect coordinated roles in immune cell function (e.g., T-cell migration, antigen presentation) and maintenance of cellular structure and signaling in healthy, young tissues. They lose this correlation in older individuals. This indicates a loss of coordinated regulation which could result in a loss of cytoskeletal regulation (PFN1) and altered adhesion (LIMD2) could impair immune responses in aging. LIMD2 shows up multiple times in Table 1 which is not surprising as we identify it as a hub node in the next section.

Table 1: Gene pairs with their correlation coefficients across different age groups. The Correlation Young, Middle, Old show the Pearson correlation coefficient for each gene pair in their respective age group.

Cell Type	Gene1	Gene2	Correlation Young	Correlation Middle	Correlation Old
central memory CD4-positive	LIMD2	PFN1	0.81	0.41	-0.12
central memory CD4-positive	LIMD2	SF1	-0.82	-0.50	-0.09
central memory CD4-positive	LIMD2	MYL12A	0.80	0.49	0.11
gamma-delta	SFPQ	RPS28	-0.81	-0.47	-0.07

4.3 Gene Network

Even though the gene pairs shown in the previous section have a very big change in correlation it might only be with one gene. To find genes that are more affected across age groups we created a network of the genes. The nodes in the networks are the genes in the correlation matrix and two nodes share an edges if the correlation between the genes changes significantly from the young to the old group (a difference is considered significant if $|z| > 2.576$) and the absolute log fold change between young and old is bigger than 2. Figure 3 shows the node degree distribution of these networks. For most of these networks the number of genes with a low degree are very high while the number of genes with a high degree are very low. This is to be expected for gene networks. A way to validate these results is to fit a power law to the node degree distribution [6][8]. Most of the networks follow this line, however the central memory CD4-positive cell type does not, the number of genes increase with the degree at the start of the plot. Even though we expect it to go down.

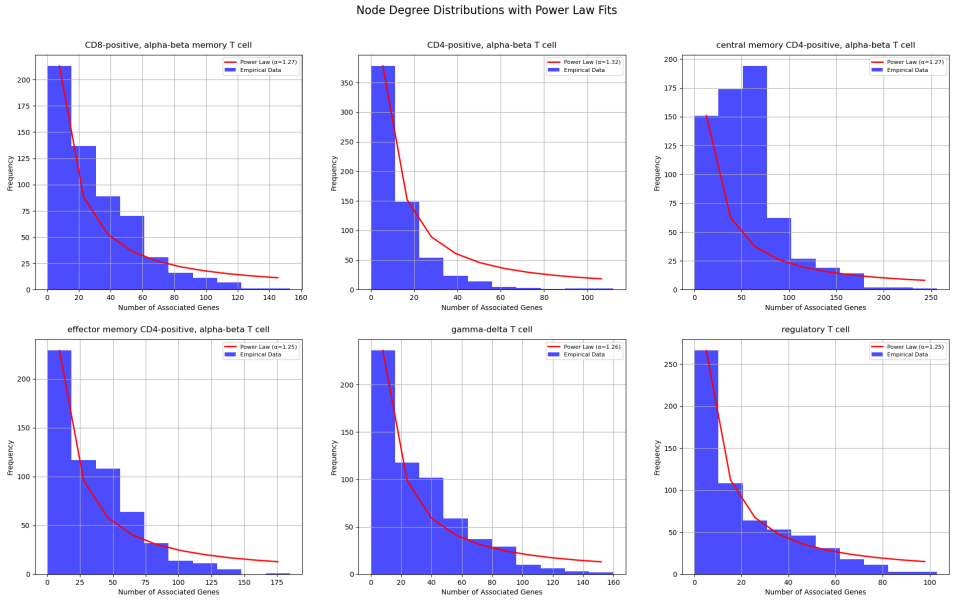

Figure 3: The Node Degree Distribution of all the gene networks one for each cell type. The x-axis shows the "Number of Associated Genes". This is the number of connections a node has, also known as the degree. And the y-axis shows the "Frequency", this is the number of genes that have this degree. The blue bars are the actual data and the red line is the fitted power law.

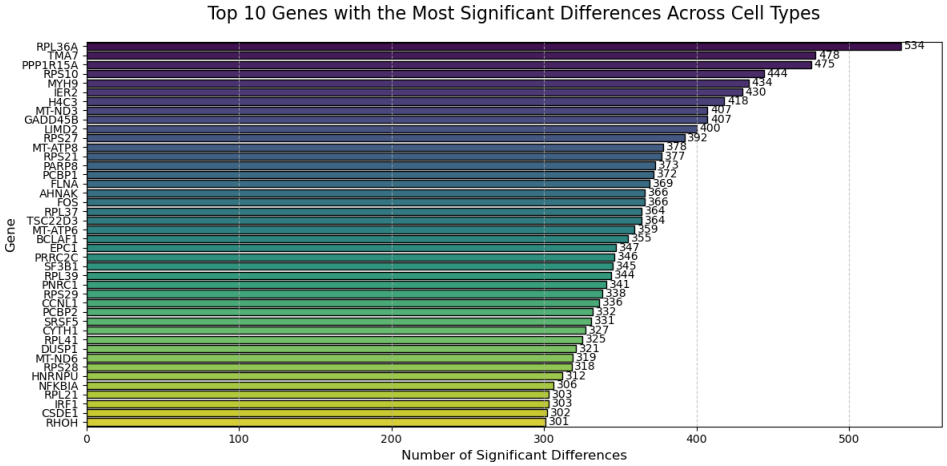

Figure 4: This figure shows genes that have a high degree across all the networks by adding those degrees together. The genes with a degree of at least 300 are shown. The x-axis shows the "Number of Significant Differences" which means how often that gene had a significant difference with another gene from the young to the old group. The y-axis shows which gene it is.

The genes with a very high degree have a lot of genes that they change correlation with from young to old. To find out which genes are the most important across cell types we added the degrees of all networks together for all cell types except the central memory CD4-positive cell type because it does not follow the power law. We then select every gene with a total degree higher than 300 (The average number of nodes in the networks is 600. 600 * 5 networks * 10% = 300). The genes that we selected are shown in Figure 4.

To better understand these genes, we used the Enrichr tool [1]. This tool shows biological processes that all or most of the genes that you give it are involved in. Figure 5 shows the biological processes that these genes are involved in. However all the processes have an adjusted p-value higher than 0.05, so none of them are significant. This means that there are a lot of different genes in the hub genes and it is not one specific biological process that stands out.

However, there are a lot of Ribosome genes like RPL41, RPS28, RPS27, RPL21, RPS29, RPL36A, RPL37, RPS10, RPL39, and RPS21, and it has been shown that ribosome function degrades with age by Kevin C Stein et al. [11]. There are also some mitochondrial genes like MT-ND3, MT-ATP8, MT-ATP6, and MT-ND6 and mitochondrial functional decline has also been linked to aging. "Mitochondrial functional decline and accrual of damaged mitochondria in various tissues is associated with aging" (Page 11, Sarika Srivastava) [10].

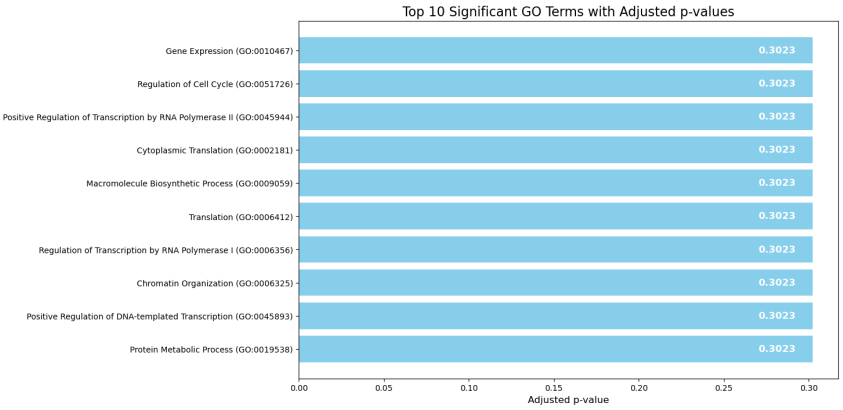

Figure 5: The results of the enriched hub genes. It shows the adjusted p-values of the hub genes entered in the Enrichr tool. The database used is GO Biological Process 2025. The name shows the biological process and the adjusted p-value shows how significant they are.

4.4 Prediction Model

Now that we have gene networks for all the cell types, we selected the hubs of the effector memory CD4-positive alpha-beta T cell network. A hub is a gene that has connections with at least 10% of the other genes. In this case that is 58 connections. We then used the hub genes to train a prediction model in the same way as Enikő Zakar-Polyák et al. trained their ElasticNet regression model [14]. Figure 6 compares our model (left) with theirs (right). Our model has a Mean Absolute Error of 9.7385 and a correlation of 0.3797, their model has a Mean Absolute Error of 9.2808 and a correlation of 0.4188. So, our model did not outperform theirs.

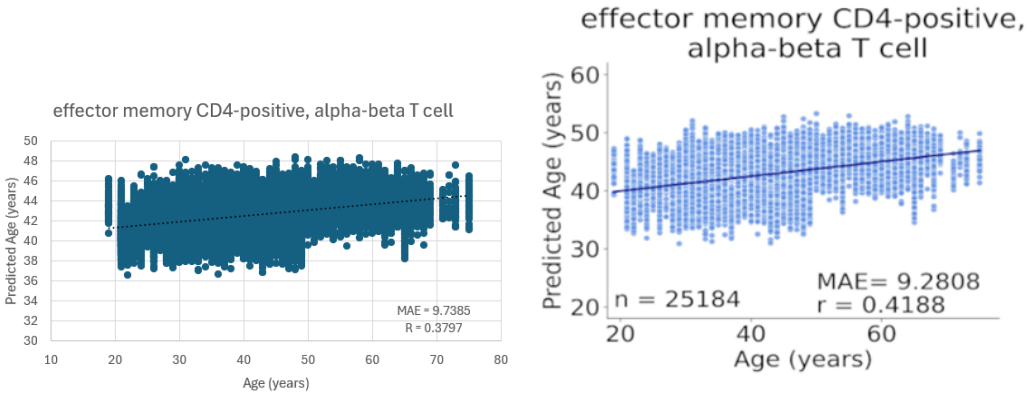

Figure 6: Left figure our prediction results of the model trained on the hub genes of effector memory CD4-positive alpha-beta T cell network, the right figure the prediction results of Enikő Zakar-Polyák et al.[14]. The x-axis shows the age in years and the y-axis the predicted age in years. MAE is the Mean Absolute Error, R is the correlation between Age and Predicted Age and n is the number of cells in the figures which is the same for left and right.

5 Discussion

5.1 Limitations and Reflection

This study had limitations that might have influenced the final results. This includes the population groups of the used data which are only from Asian countries. This reduces the generalizability to other population groups. We also only used blood cells during this research which limits generalizability to other tissues. The strict thresholds ($p < 0.01$, $|\text{LFC}| \geq 0.7$) ensured biological relevance but may have excluded other important gene pairs. The power law validation strengthened hub identification but excluded the central memory CD4-positive network from the final hubs. The threshold to define a gene as a hub is an arbitrary threshold, it could have been stricter or more lenient. Due to all of these limitations the results cannot be taken too generally. We cannot conclude that the result would be the same for different datasets with other cell types or other population groups. Or if we decided to use different thresholds.

5.2 Explanation Results

Figure 2 shows the number of gene pairs that have significant correlation differences across the three age groups for 6 cell types. Central memory CD4-positive alpha-beta T cell type shows an abnormally large number of significant pairs compared to the other cell types. 2508 compared to less than 500 for the other cell types. One possible explanation for this is that the central memory cell type also has a lot more cells in the dataset compared to the other cell types (111,905 cells compared to 39,218 cells in the next biggest subset CD8-positive alpha-beta memory T cell). The higher quantity of cells might have revealed

significant correlations between genes that would normally not be identified due to a lack of expressions in smaller subsets.

The big number of significant gene pairs for the central memory cell type might also explain why its node degree distribution does not follow a power law (see Figure 3). The big number of significant pairs means that there are more genes that are connected. This results in more genes with around 50 connections than genes with 0 to 10 connections. But after the small increase at the start of the node degree distribution it does drop very fast.

When we used the hub genes of the effector memory cell type to train a ElasticNet linear regression model (see Figure 6), it did not outperform the model made by of Enikő Zakar-Polyák et al.[14]. One possible reason it did not improve the performance is the inherent limitations of ElasticNet regression in capturing non-linear relationships. While ElasticNet effectively handles multicollinearity and performs feature selection through its L1/L2 regularization, it may fail to model complex, higher-order interactions between the hub genes. Additionally, our correlation-based hub gene selection prioritized coordinated expression changes rather than individual predictive power, potentially excluding genes with strong linear age associations that ElasticNet could readily exploit. This suggests that alternative modeling approaches (e.g., graph neural networks or non-linear regression) might better leverage the network topology information captured by our correlation analysis.

6 Conclusions and Future Work

This study investigated whether analyzing changes in gene-gene correlations across age groups could improve biological age prediction models. Through computational analysis of 1,058,909 blood cells from 508 donors, we found significant age-dependent correlation changes between gene pairs, one example is LIMD2-PFN1 (from $r = 0.81$ in young to $r = -0.12$ in old). Multiple hub genes are identified showing consistent correlation changes across multiple cell types. And we demonstrated that hub genes can predict age with comparable accuracy (MAE=9.74 years) to existing models. So there clearly are differences between correlations of genes in young and old individuals. But the identified genes have not been able to improve the existing age prediction model.

For future work we could perform the same experiments on different datasets with other population groups or tissues to see if the results are similar or very different. The number of age groups could also be increased, this could provide more insight in how fast the correlation changes from age to age. Other types of machine learning models could be trained on the found hub genes like an artificial neural network for example. This might find patterns that cannot be found by linear regression models. A more specific example would be to train an artificial neural network on the number of expressions for the hub genes in the blood cells. The neural network would then classify the blood cell in one of the three age groups.

New and unanswered questions that resulted from this research are: Do the observed correlation patterns generalize across ethnic groups beyond Asian populations? Do other tissues (e.g., skin, muscle, neurons) show similar age-related correlation changes? Does using the hub genes as features in other machine learning techniques improve the accuracy of age prediction/classification?

References

[1] Edward Y. Chen. Enrichr — maayanlab.cloud. https://maayanlab.cloud/Enrichr/, 2013.

[2] Creative Commons. Deed - Attribution 4.0 International - Creative Commons — creativecommons.org. https://creativecommons.org/licenses/by/4.0/.

[3] GeneCards Human Gene Database. LIMD2 Gene - GeneCards | LIMD2 Protein | LIMD2 Antibody — genecards.org. https://www.genecards.org/cgi-bin/carddisp.pl?gene=LIMD2.

[4] John Elflein. Influenza deaths by age group U.S.| Statista — statista.com. https://www.statista.com/statistics/1127698/influenza-us-deaths-by-age-group/, 2024.

[5] Olga Ibañez-Solé, Irantzu Barrio, and Ander Izeta. Age or lifestyle-induced accumulation of genotoxicity is associated with a length-dependent decrease in gene expression. *iScience*, 26(4):106368, April 2023.

[6] Ryan S. McClure, Christopher C. Overall, Jason E. Mcdermott, Eric A. Hill, Lye Meng Markillie, Lee Ann Mccue, Ronald C. Taylor, Marcus Ludwig, Donald A. Bryant, and Alexander S. Beliaev. Network analysis of transcriptomics expands regulatory landscapes in synechococcus sp. pcc 7002. *Nucleic Acids Research*, 44(18):8810–8825, 2016.

[7] Elena Ratti and James D. Berry. Chapter 42 - amyotrophic lateral sclerosis 1 andÂ many diseases. In Thomas Lehner, Bruce L. Miller, and Matthew W. State, editors, *Genomics, Circuits, and Pathways in Clinical Neuropsychiatry*, pages 685–712. Academic Press, San Diego, 2016.

[8] Matthias Scholz. Network science: node degree distribution — network-science.org. http://www.network-science.org/powerlaw_scalefree_node_degree_distribution.html.

[9] Sourena Soheili-Nezhad, Olga Ibáñez-Solé, Ander Izeta, Jan H J Hoeijmakers, and Thomas Stoeger. Time is ticking faster for long genes in aging. *Trends Genet.*, 40(4):299–312, April 2024.

[10] Sarika Srivastava. The mitochondrial basis of aging and age-related disorders. *Genes*, 8(12), 2017.

[11] Kevin C. Stein, FabiÃ¡n Morales-Polanco, Joris van der Lienden, T. Kelly Rainbolt, and Judith Frydman. Ageing exacerbates ribosome pausing to disrupt cotranslational proteostasis. *Nature*, 601(7894):637–642, 2022.

[12] P. Tao, X. Chen, L. Xu, J. Chen, Q. Nie, M. Xu, and J. Feng. Limd2 is the signature of cell aging-immune/inflammation in acute myocardial infarction. *Current Medicinal Chemistry*, 31(17):2400–2413, 2024.

[13] Suzanne Wakim and Mandeep Grewal. 23.8: Adulthood — bio.libretexts.org. https://bio.libretexts.org/Bookshelves/Human_Biology/Human_Biology_(Wakim_and_Grewal)/23%3A_Human_Growth_and_Development/23.8%3A_Adulthood, 2021.

[14] Enikő Zakar-Polyák, Attila Csordas, Róbert Pálovics, and Csaba Kerepesi. Profiling the transcriptomic age of single-cells in humans. *Commun. Biol.*, 7(1):1397, October 2024.

A Appendix

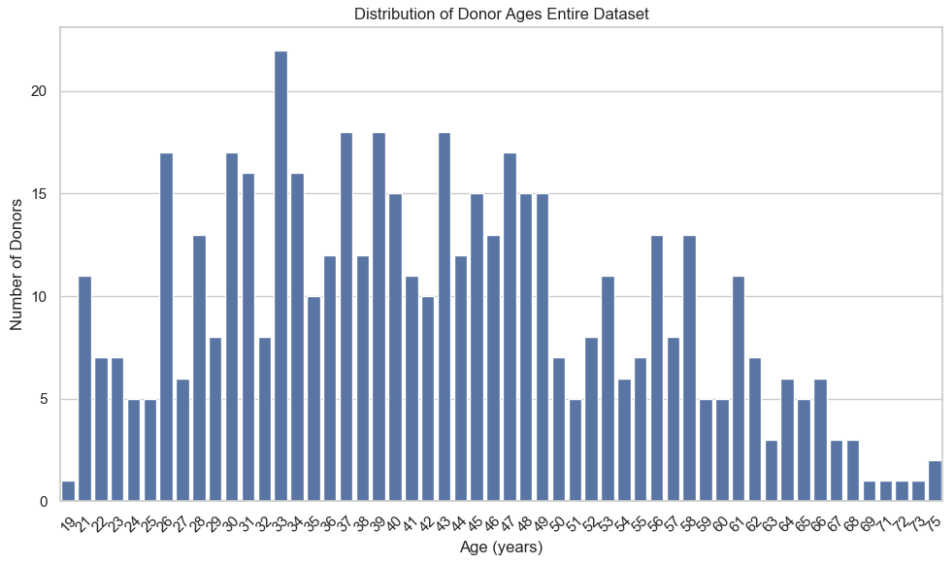

Figure 7: The distribution of the ages for the 508 donors in the dataset. The x-axis shows the age of the donors and the y-axis shows the number of donors in the dataset that have this age. For example there is only one 19 year old donor and there are no donors older than 75 or younger than 19.

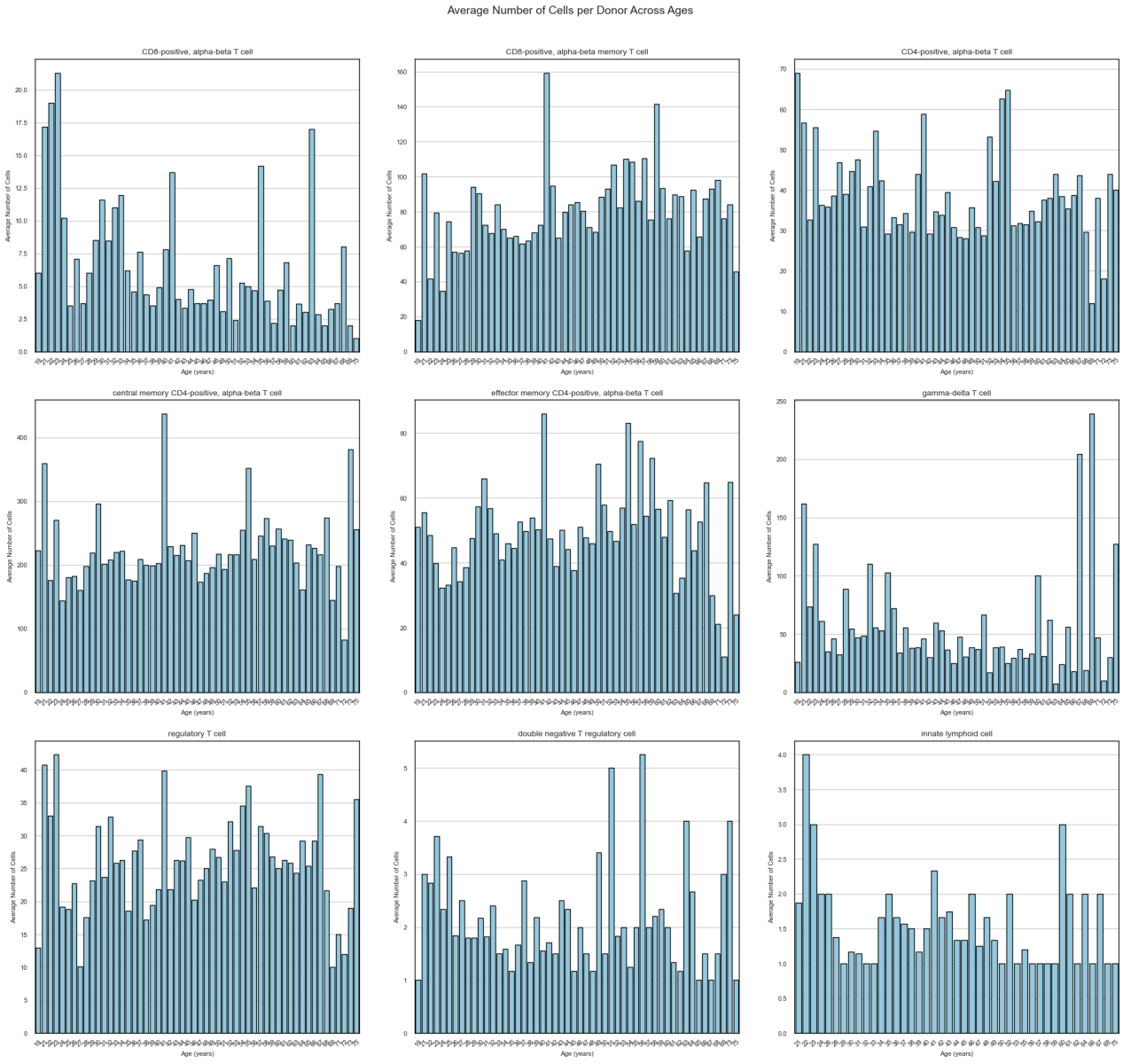

Figure 8: The average number of cells for every age for every cell type. The x-axis shows the age (in years) of the donors and the y-axis shows the average number of cells that are in the database for the specific cell type of that plot.