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

Author: Tycho Grapendaal (T.Grapendaal@student.tudelft.nl)

Supervisors: Marcel Reinders, Bram Pronk, Inez den Hond, Gerard Boulandl

1. Introduction

There are two types of ages. Chronological age: the time since birth. And biological age: the age of our cells. The difference between these two ages can tell us if someone is healthy or not. Right now there exist models that can predict someone's biological age based on their gene expressions. However these models can still be improved. To find new ways to improve these models, we looked at correlations between genes from different blood cells.

Correlations between genes can tell us if genes are expressed together, and if the biological processes they are responsible for work together. We want to know if these correlations change with age and if we can use this information to better predict someone's biological age.

2. Research 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?

3. Method

Preproccessing

- Create a subset of the data based on cell type
- Remove genes with a low expression count
- Taking the average for all the cells of the same donor
- Remove donors with less than 10% of the median cells
- Create 3 subsets: young (19-30), middle (40-50), old (60-75)

Fisher's z-transformation

- Comparing two correlation coefficients in different groups
- If $|Z| > 2.576$ there is a p value of 0.01

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) z = \frac{z_{\text{diff}}}{SE} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

Correlation Analysis

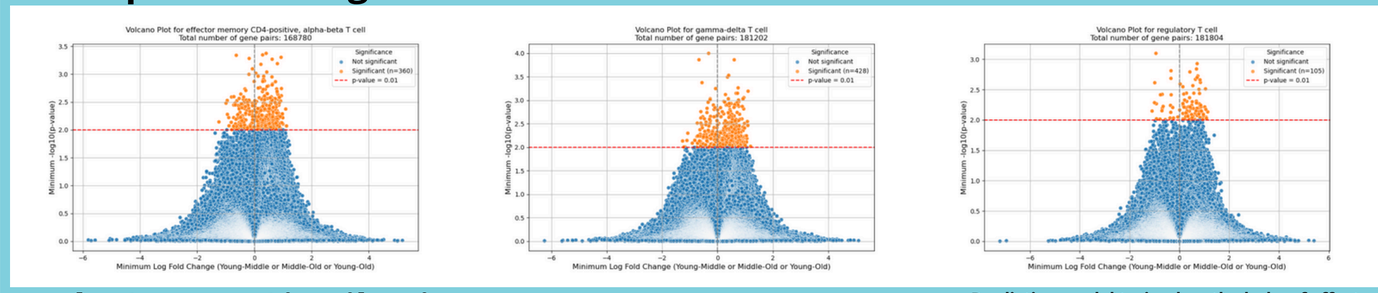
- Calculate the Pearson correlations for gene pairs across the young (19-30), middle (40-50), old (60-75) groups
- Create a network of all the genes based on a significant difference in correlation. Two genes are connected in the network if there is a significant difference in correlation from young to the old

Prediction Model

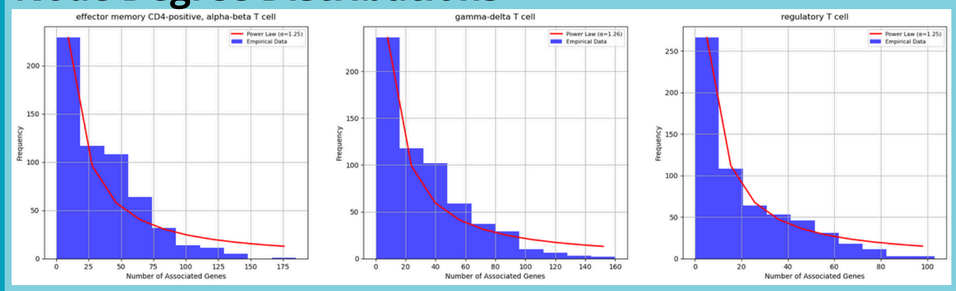
- Identify the hub genes by filtering for genes that have a connection with at least 10% of the other genes
- Use the hub genes of a specific cell type as features to train a ElasticNet linear regression model

4. Results

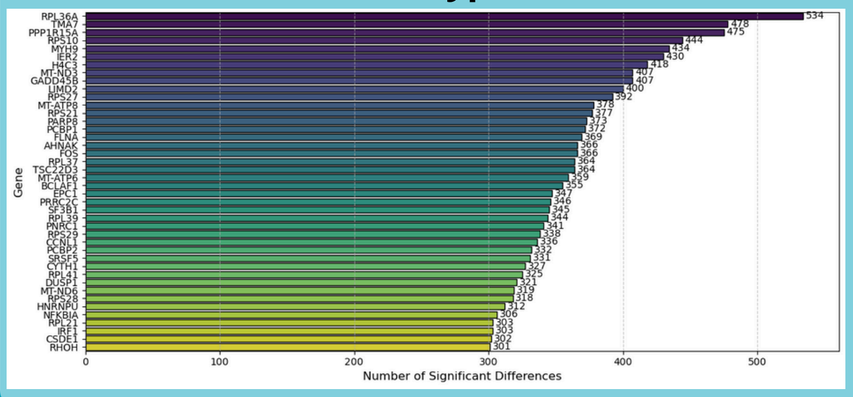
Gene pairs with significant correlation differences



Node Degree Distributions

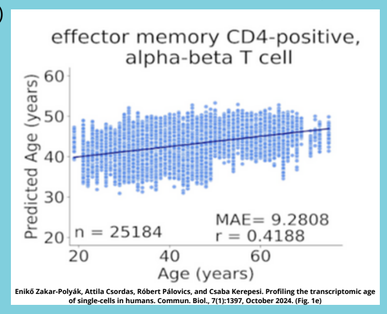


Hub Genes Across Cell Types

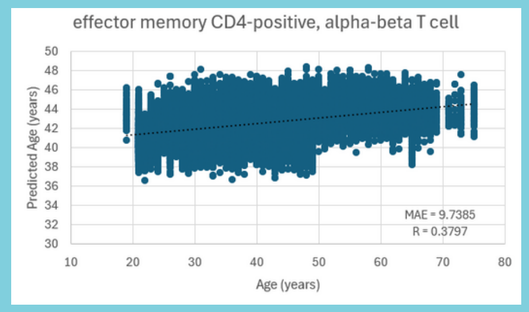


Prediction model trained on the hubs of effector cell type

Our model (bottom) has a Mean Absolute Error (MAE) of 9.7385 and a Correlation between age and predicted age (R) of 0.3797. This did not outperform the model of Enikő Zakar-Polyák et al. (top)



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. (Fig. 1e)



5. Conclusion

- There are significant differences between features in young and old individuals. But it did not improve current age prediction models.

6. Future Work

- Training other machine learning models like neural networks
- Performing Correlation Analysis on different cells like skin, muscle or neurons