Can We Use Physical Characteristics of Genes to Predict Age-Related Changes in Expression?

[CSE3000 Research Project] Lovro Mlikotić
(L.Mlikotic@student.tudelft.nl)

TU Delft

[1] Introduction

This research aims to deepen our understanding of aging at the molecular level and support future strategies for detecting, mitigating, or potentially reversing age-related gene expression changes. To support this goal, the study investigates whether physical gene-level characteristics, such as gene length, GC content, and chromosomal location, can predict age-related changes in gene expression. It specifically asks whether a set of intrinsic gene features can be utilized to predict whether a gene is more likely to be differentially expressed in young versus old individuals. To explore this, the study combines machine learning classifiers with statistical analysis of gene features evaluate their predictive power.

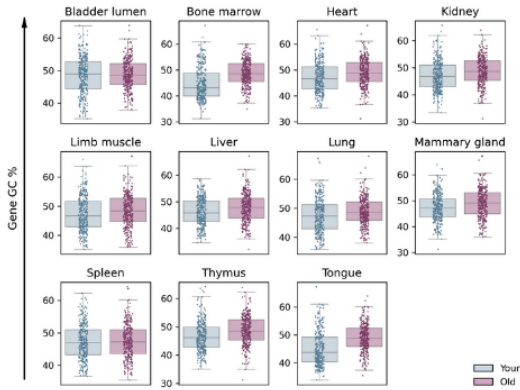
[2] Data & Processing

This study is conducted on the Tabula Maris Senis (TMS) dataset, a large-scale dataset, containing gene expression profiles from 245,389 individual cells and 78,258 genes.

The dataset was cleaned and processed, with sample ages (in months) grouped into "young" and "old" to enable differential expression analysis: 1m, 3m → young, 18m, 21m, 24m, 30m → old

[3] Characteristic Analysis

Gene length was also considered, but since it is a studied characteristic already suggested to influence age-related expression patterns [1][2], it was not analyzed on it's own in this research.

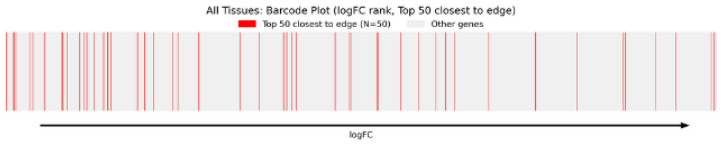


Each panel shows the GC content distribution of the 300 most differentially expressed genes overexpressed in young (blue) and old (purple) samples for a given tissue. Boxplots display summary statistics, while the overlaid dots correspond to individual genes.

Next, GC content (defined as the proportion of guanine and cytosine bases in a gene) was analyzed for its association with age-related expression changes. Across most tissues, genes overexpressed in old mice showed higher GC content compared to those overexpressed in young mice. This trend was statistically significant in the majority of tissues, with the strongest effects in bone marrow, tongue, and thymus ($p < 10^{-8}$). Only bladder lumen and spleen showed no significant difference. These results suggest a broad link between higher GC content and increased gene expression in older tissues.

Last, the distance of a gene from the ends of its chromosome (defined as the shorter distance to either chromosomal end) was analyzed. Among significantly differentially expressed genes ($p < 0.05$), those closest to chromosome ends tend to be more highly expressed in young mice. This effect is strongest for the very closest genes, suggesting a possible link between chromosomal position and age-related expression changes.

Each vertical line represents a gene. The genes are ranked by their logFC, with the lowest negative logFC being on the left, and the highest positive logFC being on the right. The vertical lines colored in red represent the 100 genes closest to the edge of their chromosome.



[4] DEG Classes

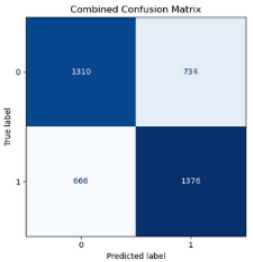
The dataset was pruned of all genes with a non-statistically significant differential expression between young and old mice (p-value < 0.05). The genes were then assigned one of the two class labels:

- the 20% most differentially expressed genes in young mice → class 0
- the 20% most differentially expressed genes in old mice → class 1

The middle 60% were discarded.

[5] Classifier Results

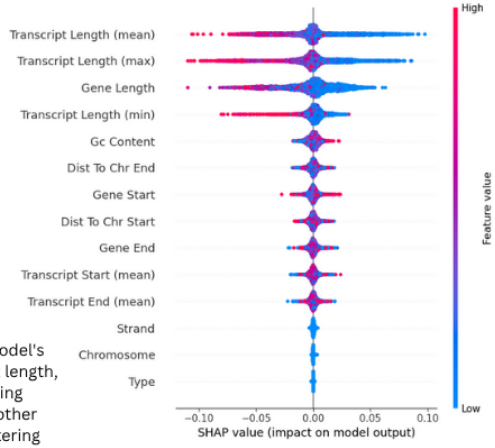
Multiple classifiers with varying complexity, along with different data labeling strategies, were tested. Regression-based methods were also explored but failed to produce interpretable results. The best-performing models, XGBoost and Random Forest, both achieved modest accuracy of around 66%, with parameter tuning providing minimal improvement. Given their nearly identical performance, feature importance rankings, and SHAP plots, further analysis focuses on the Random Forest results.



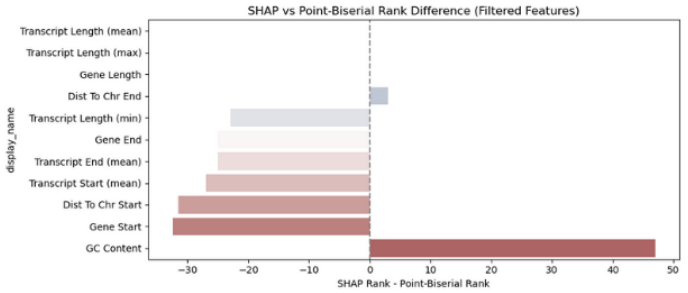
Random Forest. All Tissues. Combined Confusion Matrix.

Classifier	Class	Precision	Recall	F1-score	Support
XGBoost	0	0.6598	0.6605	0.6601	2044
	1	0.6598	0.6592	0.6595	2042
	Avg	**0.6598**	**0.6598**	**0.6598**	**4086**
Random Forest	0	0.6630	0.6409	0.6517	2044
	1	0.6521	0.6738	0.6628	2042
	Avg	**0.6575**	**0.6574**	**0.6573**	**4086**

SHAP was used to interpret feature contributions to the model's predictions. Gene length and related features, like transcript length, emerged as the most influential, with lower values pushing predictions toward the overexpressed-in-old class. Most other features showed minimal and inconsistent influence, clustering near the center of the SHAP beeswarm plot.



Random Forest. All Tissues. Combined SHAP Beeswarm. Each point represents an observation (gene). The color indicates the magnitude of the feature value for that observation. The horizontal position of the dot indicates the SHAP value. If it is to the right, it contributes toward class 1, if it is to the left, it contributes toward class 0. The magnitude of the horizontal position signifies contribution strength.



The plot shows the difference in feature importance rankings between SHAP and PB. The features on the plot are sorted by the absolute value of this difference. Bars extending right indicate features ranked higher by PB than by SHAP. The bars extending left indicate the opposite. The blue-to-red color gradient is solely a visual aid, and only signifies the absolute difference between the SHAP and PB ranks.

To validate and better understand the key features identified by the classifier, a point-biserial correlation analysis was performed. This method measures the relationship between continuous gene features and the binary expression label (upregulated in young vs. old). The results showed strong agreement between the point-biserial analysis and SHAP values from the classifier, with both approaches ranking mean transcript length, maximum transcript length, and gene length as the top three predictors, in the same order. GC content ranked notably higher in the point-biserial analysis than in the SHAP rankings, highlighting some differences. Overall, both methods produced consistent feature importance rankings and agreed on the direction and influence of key features driving age-related gene expression changes.

[6] Conclusion

The two most meaningful conclusions are these:

1. Across all experiments, classifier performance plateaued at around 66% accuracy, even after trying various models and tuning parameters, suggesting that the current feature set lacks the predictive power needed for stronger results. Consistent overlap between classifier-derived SHAP rankings and independent point-biserial analyses further supports that the classifiers have likely extracted all available signal from the current data.
2. Among all features, gene length stood out as the most consistent and meaningful predictor, with longer genes tending to be overexpressed in young mice and shorter genes in older mice. This aligns with prior research and highlights gene length as a key feature for understanding age-related gene expression patterns, even though, on it's own, it does not have enough predictive value to train a high-accuracy classifier.

[7] Future Work

The 10-week deadline limited the scope of this research, leaving several important directions for future work. A logical order for pursuing these next steps would be:

- repeat the study using transcript-level rather than gene-level data
- use dimensionality reduction methods like t-SNE and UMAP to evaluate whether the predictive limit of the current feature set has been reached
- extend the feature set to include additional biologically relevant characteristics

supervisors: I. den Honed, B. Pronk, G. Bouland, M. Reinders

[1] Ibañez-Solé O., Barrio I., Izeta A. (2023). Length-dependent decrease in gene expression linked to age and genotoxic stress. iScience, 26(4), 106368.
[2] Soheili-Nezhad S., Ibañez-Solé O., Izeta A., Hoeijmakers J.H.J., Stoeger T. (2024). Long genes show accelerated transcriptional decline with aging. Trends Genet., 40(4), 299–307.