Feasibility of Tissue-specific Gene Expression as a Predictor of Subject Age

Anthony Toloczko, Luke Schoen

Introduction:

Understanding how gene expression changes with age is critical for understanding the biological processes that underlie human development, aging, and age-related diseases. Recent advancements in transcriptomic technologies, and the availability of large-scale databases of gene expression information have enabled more in-depth investigations into the relationship between gene activity and chronological age. In this study, we analyze gene expression profiles obtained from the Genotype-Tissue Expression (GTEx) portal, which is a comprehensive resource that provides RNA sequencing data across multiple tissues from the human body. By examining how gene expression levels vary across individuals of different ages, we aim to identify tissue-specific genes that are associated with aging.

Methodology:

In order to assess the correlation between gene expression and subject age, three files were obtained from the GTEx portal. These files include the GTEx Analysis V10, a bulk tissue expression dataset containing gene expression counts for each sample from each subject, and the V10 analysis annotations related to Subject Phenotypes and Sample Attributes. All analysis performed on these datasets was performed in RStudio.

The datasets were read into RStudio using the Tidyverse and dplyr packages, and then wrangling was performed on the raw gene expression dataset to add subject IDs (SUBJID) and age brackets (AGE) from the Subject Phenotypes dataset, as well as general tissue sample source (SMTS) and specific tissue sample source (SMTSD) from the Sample Attributes dataset. Subject ages have been generalized to age brackets of 10 years, to protect subject information. Once the information was read in and cleaned up, several methods of regression were used to test the correlation between tissue-specific gene expression and subject age. First, linear regression was applied to the modified gene expression dataset, to examine any particularly strong correlated genes. The data set was then subset by specific tissue of origin. A regression analysis was then preformed to check for correlation. Finally, principal component analysis (PCA) was performed, and another regression analysis ran. A subset of subjects was then used to predict subject age based on gene expression using both a K-Nearest Neighbors (KNN) algorithm and traditional regression analysis.

There was also an attempt made to create a linear regression in a similar fashion as was previously done, although with the goal of modeling correlation between tissue-type expression and subject age. This, however, was not nearly as successful. The model did find significant age correlation across sixteen of the tissue types in the overall dataset, although the correlation proves to be unreliable due to the incredibly insignificant R^2 values which leads us to believe that the model is not well fit at all, and the correlation could be false, thus no predictions were made with this model.[4][5]

Results:

The regressions performed on this dataset were performed in a variety of ways, to test the best methodology for regressing gene expression against age. First, a regression was run on each gene in each tissue category (i.e. the Cerebral Cortex, Cerebral Hemisphere, etc., were grouped together under ‘Brain’). Each gene had an extremely small R^2 value, the greatest of which was

.105 for the gene ENSG0000172667.11 in the group Adipose Tissue. Associated with this was an extremely small P-value. Generally, the tissues with the highest R^2 values were Blood and Adipose tissue. These tissues also had the most available samples. P-values ranged from near 0 to almost 1. Generally, no gene by itself in any group was a particularly impressive predictor of subject age based on the expression of that gene in relation to the tissue the sample was taken from. The correlation was modest at best and explained little of the variance between samples. This may seem like an obvious result, but it was worth examining simply to rule out the possibility.

The samples were then grouped together by specific tissue of origin for the sample. Principal component analysis was then conducted to reduce the dimensionality of the dataset for each group. A linear regression was then performed to evaluate the relationship between the expression of those genes which most explain the variance in relation to age for each specific tissue.[1]

In this method of exploring the relationship between gene expression and subject age, the correlation statistics were generally improved over those of the by-gene-by-tissue method. The majority of correlation values were between .05 and 1.5, with much improved P-values. These improved correlation statistics still suggest that there is only a moderate relationship between tissue-specific gene expression and subject age.[2]

Given the improved correlation statistics, a KNN algorithm was then used to split the samples for each tissue into training and testing groups, and predictions for the ages of the test subjects were then made and Mean-Squares Error (MSE) collected for these predictions. KNN MSE ranged between 72.94 and 380, with a mean value of 153.97. The number of samples available for each specific tissue type varied widely. The Portal Tract of the liver only had two expression samples available, while Skeletal Muscle had 818 expression samples available. The KNN predictions were often wildly inaccurate, suggesting that tissue-specific gene expression is not a good basis for predicting the chronological age of a subject.[3]

Conclusions:

In conclusion, this study examined the feasibility of using tissue-specific gene expression as a predictor of subject age. Our findings suggest that while there is a modest correlation between the expression of genes and the age of the subject across the tissues included in the dataset, the predictive power of this relationship is extremely limited. The use of K-Nearest Neighbors for the prediction of a subject's age yielded low accuracy. This is likely due to the limited number of samples available for some tissues, and the inherently weak relationship between age and gene expression.

The results indicate that, although gene expression may contain some age-related information, it is not sufficient on its own to reliably predict the chronological age of a subject. Future research may benefit from integration of additional molecular features and more well-rounded sample counts. Moreover, gene expression is influenced by a variety of complex factors, such as the diverse nature of human genetics and environmental factors like pollutants, which can significantly alter gene expression. It seems to be that, while there is some slight correlation between the gene expression and age of a subject, gene expression on its own will not be a sufficient pathway to gaining insight into the biological process of aging.

Appendix:

Tissue	Linear Regression R^2	Linear Regression P-value	KNN MSE
Adipose - Subcutaneous	0.1287	1.79e-16	183.49
Adipose - Visceral (Omentum)	0.2126	8.12e-25	127.93
Adrenal Gland	0.0699	2.21e-02	221.96
Artery - Aorta	0.2679	4.07e-26	118.48
Artery - Coronary	0.1127	5.44e-04	151.92
Artery - Tibial	0.2809	7.36e-43	140.38
Bladder	0.2292	5.19e-02	193.71
Brain - Amygdala	0.2384	8.10e-07	96.56
Brain - Anterior cingulate cortex (BA24)	0.1194	1.37e-03	96.09
Brain - Caudate (basal ganglia)	0.0918	1.67e-03	96.34
Brain - Cerebellar Hemisphere	0.0672	4.31e-02	146.00
Brain - Cerebellum	0.1464	1.14e-05	89.15
Brain - Cortex	0.1663	6.65e-07	77.15

Figure 1: A Sample of Correlation Statistics from Tissue-Specific Regression

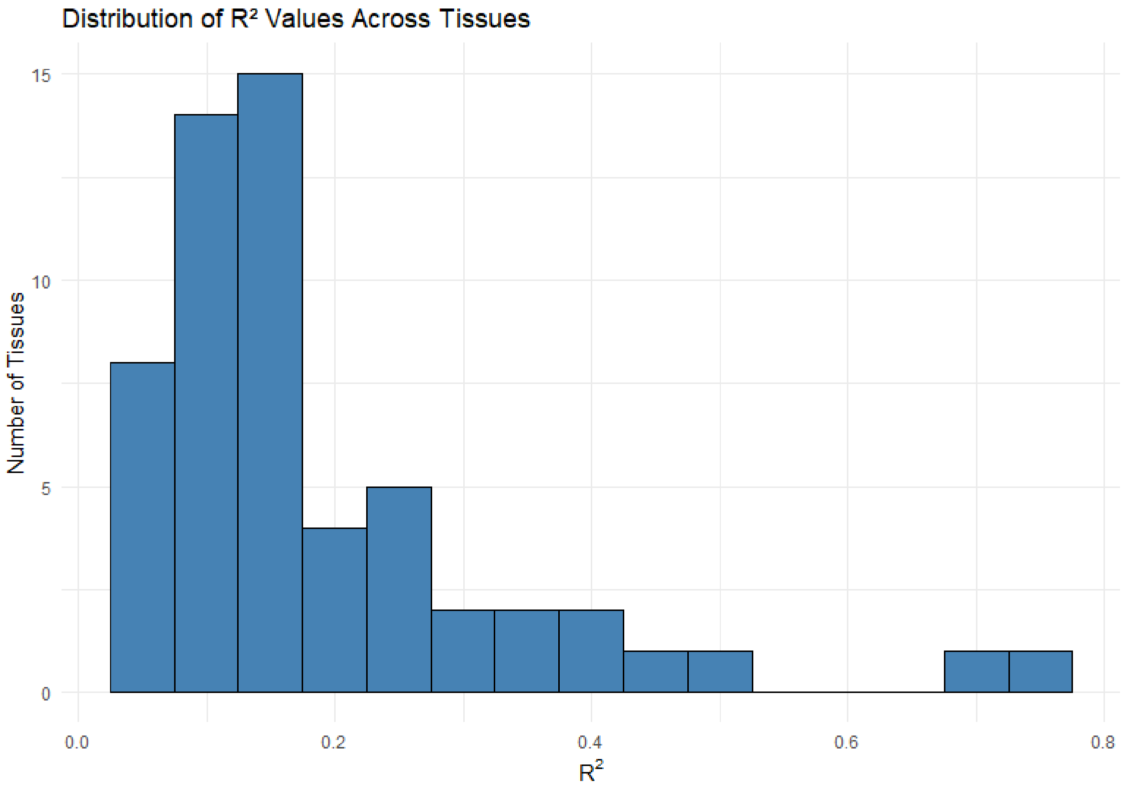


Figure 2: Distribution of Correlation Values Across Tissues

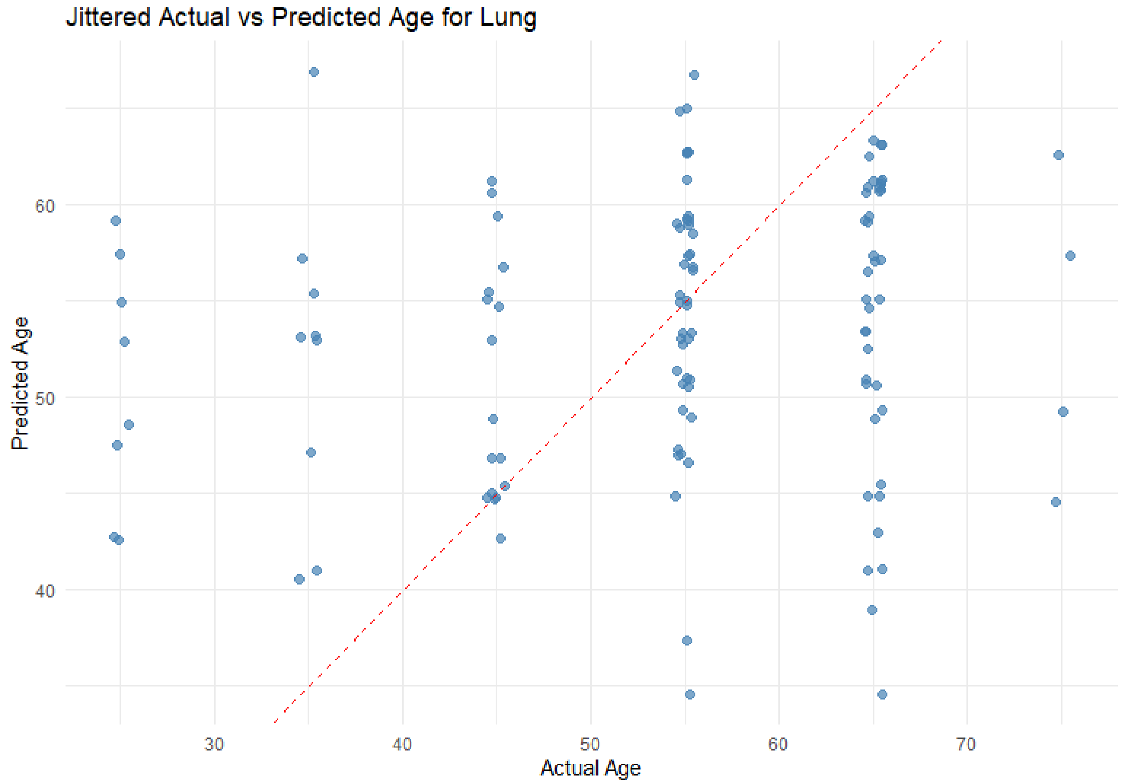
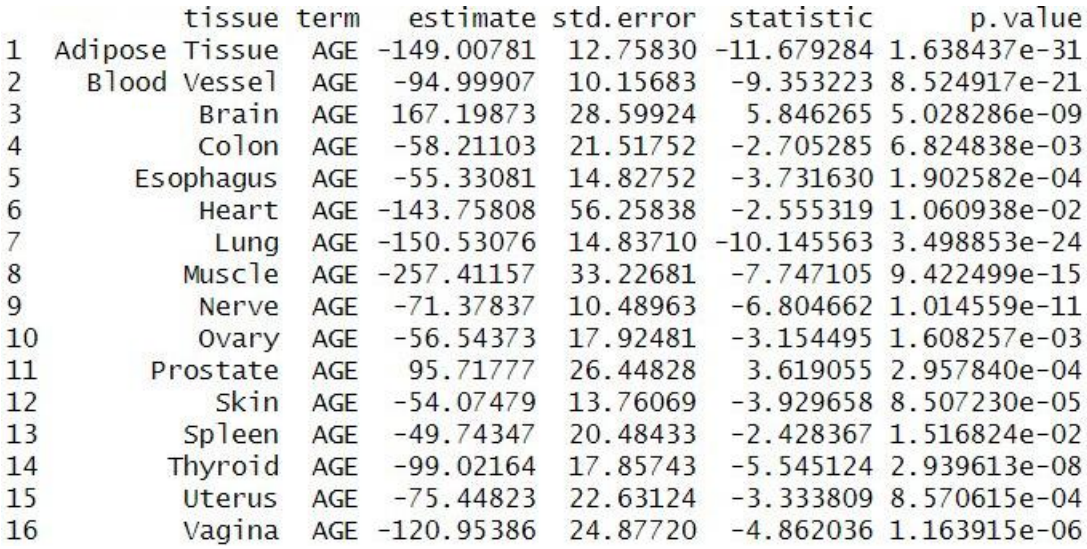


Figure 3: KNN Predicted vs. Actual Age: Lung Tissue



```
           tissue term    estimate std.error  statistic      p.value
1  Adipose Tissue  AGE -149.00781  12.75830 -11.679284 1.638437e-31
2    Blood Vessel  AGE  -94.99907  10.15683  -9.353223 8.524917e-21
3           Brain  AGE  167.19873  28.59924   5.846265 5.028286e-09
4           Colon  AGE  -58.21103  21.51752  -2.705285 6.824838e-03
5       Esophagus  AGE  -55.33081  14.82752  -3.731630 1.902582e-04
6           Heart  AGE -143.75808  56.25838  -2.555319 1.060938e-02
7            Lung  AGE -150.53076  14.83710 -10.145563 3.498853e-24
8          Muscle  AGE -257.41157  33.22681  -7.747105 9.422499e-15
9           Nerve  AGE  -71.37837  10.48963  -6.804662 1.014559e-11
10          Ovary  AGE  -56.54373  17.92481  -3.154495 1.608257e-03
11       Prostate  AGE   95.71777  26.44828   3.619055 2.957840e-04
12           Skin  AGE  -54.07479  13.76069  -3.929658 8.507230e-05
13         Spleen  AGE  -49.74347  20.48433  -2.428367 1.516824e-02
14        Thyroid  AGE  -99.02164  17.85743  -5.545124 2.939613e-08
15         Uterus  AGE  -75.44823  22.63124  -3.333809 8.570615e-04
16         Vagina  AGE -120.95386  24.87720  -4.862036 1.163915e-06
```

Figure 4: Tissue-type linear regression's sixteen significant age correlated tissue types

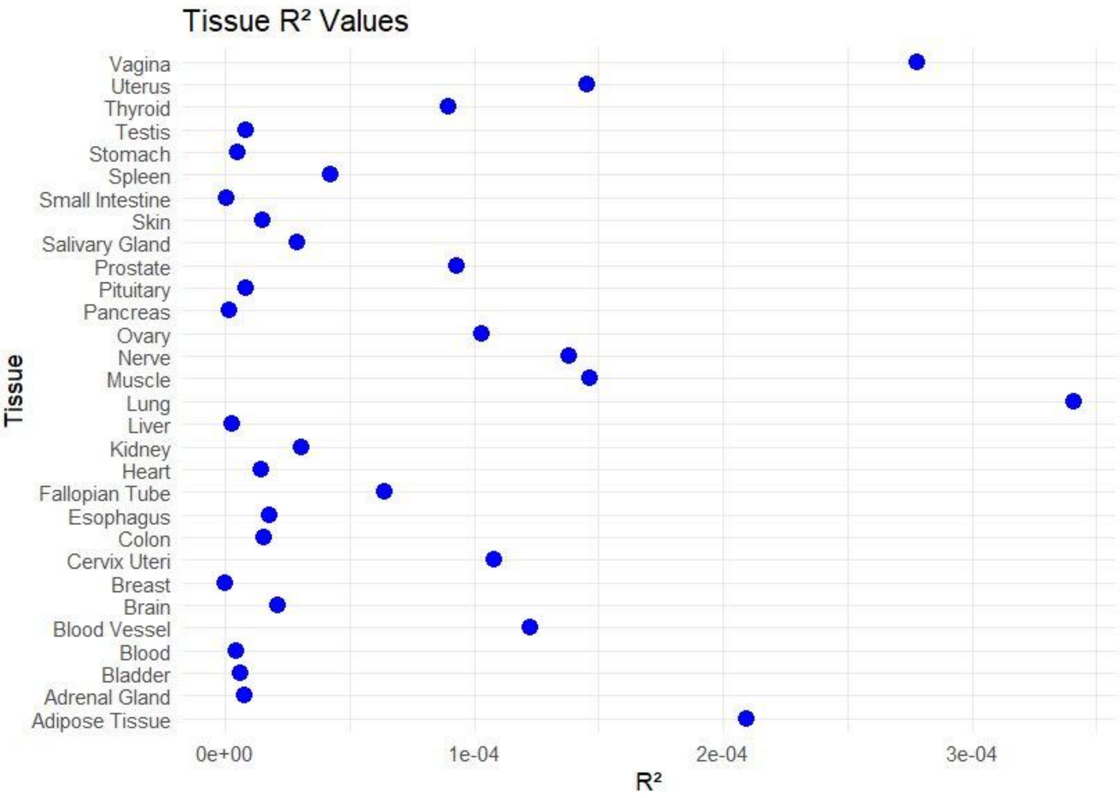


Figure 5: Tissue-type linear regression R^2 values

Sources:

The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on 04/25/2025 and/or dbGaP accession number phs000424.vN.pN on 04/25/2025.