Feasibility of Tissue-Specific Gene Expression as a Predictor of Subject Age:

An Examination of Raw Expression Counts as a Method of Determining the Age of a Subject

A. Toloczko
L. Schoen

Overview:

- Motivation
- Methods
- Results
- Conclusions
- Sources

Motivation

- Understanding how gene expression changes with age is critical for understanding the biological processes that underlie human development, aging, and age-related diseases
- Availability of large-scale databases of gene expression information have enabled more in-depth investigations into the relationship between gene activity and age
- In this study, we examine how gene expression levels vary across individuals of different ages, attempting to identify tissue-specific genes that can be used to predict the age of a subject

Methods

- Using RStudio, we loaded and merged raw gene expression data, subject phenotype data, and sample metadata into one large data frame.
- The data frame was then used to preform regression analysis.
- PCA was used to reduce dimensionality.
- A K-Nearest Neighbors algorithm was used to make predictions about a subjects age.

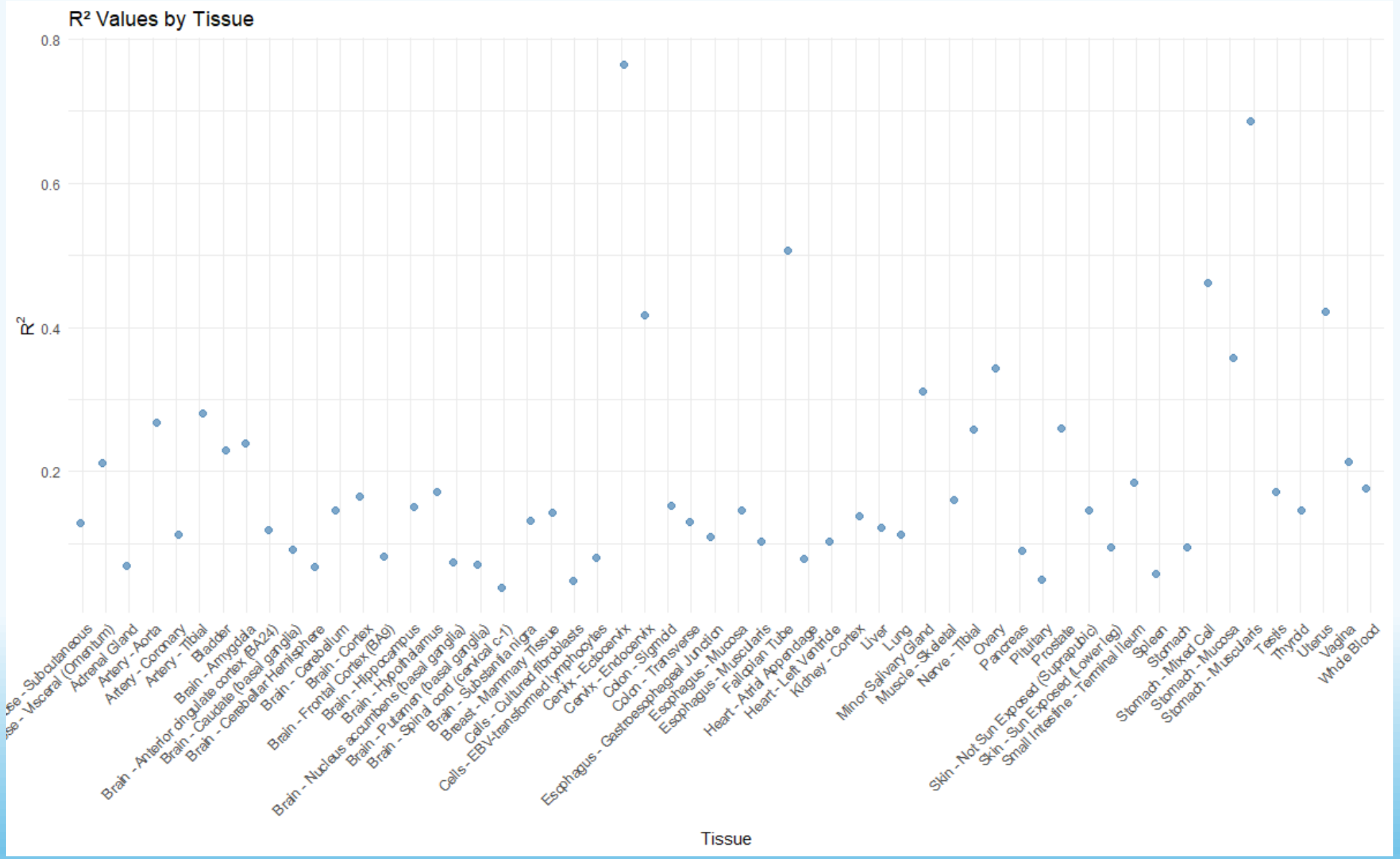
Results

- R^2 values tended to be between .1 and .3, with extremely small P-values
- Small p-values suggest that the relationship between tissue-specific gene expression and age is statistically significant, despite modesty of explained variance
- Tissues with higher R^2 values may have more stable expressions over time, or greater heterogeneity

Sample of Regression Statistics

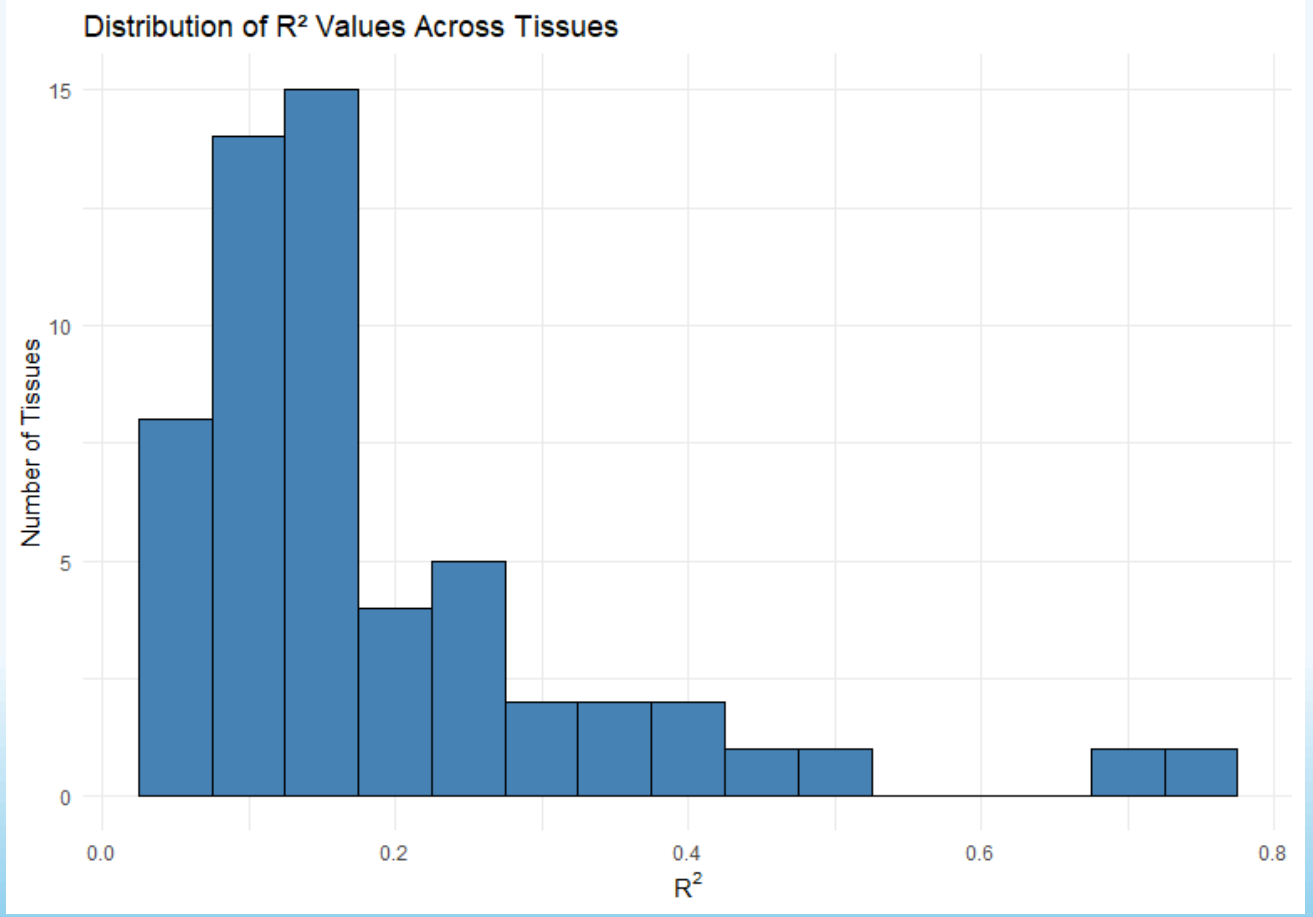
- High MSE an issue across most tissues
- Genetic diversity among humans likely root cause of high MSE
- Number of tissue samples available greatly impacts KNN prediction

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
Brain - Frontal Cortex (BA9)	0.0829	1.20e-02	77.23
Brain - Hippocampus	0.1515	1.25e-05	76.00
Brain - Hypothalamus	0.1721	8.68e-07	72.94


R² Values by Tissue
0.8
0.6
0.4
0.2
R²
se - Subcutaneous
se - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - Cultured fibroblasts
Cells - EBV-transformed lymphocytes
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Stomach - Mixed Cell
Stomach - Mucosa
Stomach - Muscularis
Testis
Thyroid
Uterus
Vagina
Whole Blood
Tissue

Distribution of R^2 Values

- Most tissue expressions have modest R^2 values, indicating they explain moderate variance
- Discrepancy between number of samples per tissue an issue in study reliability
- Environmental impacts on gene expression, and genetic diversity may cause lack of reliability



Conduction of Analysis

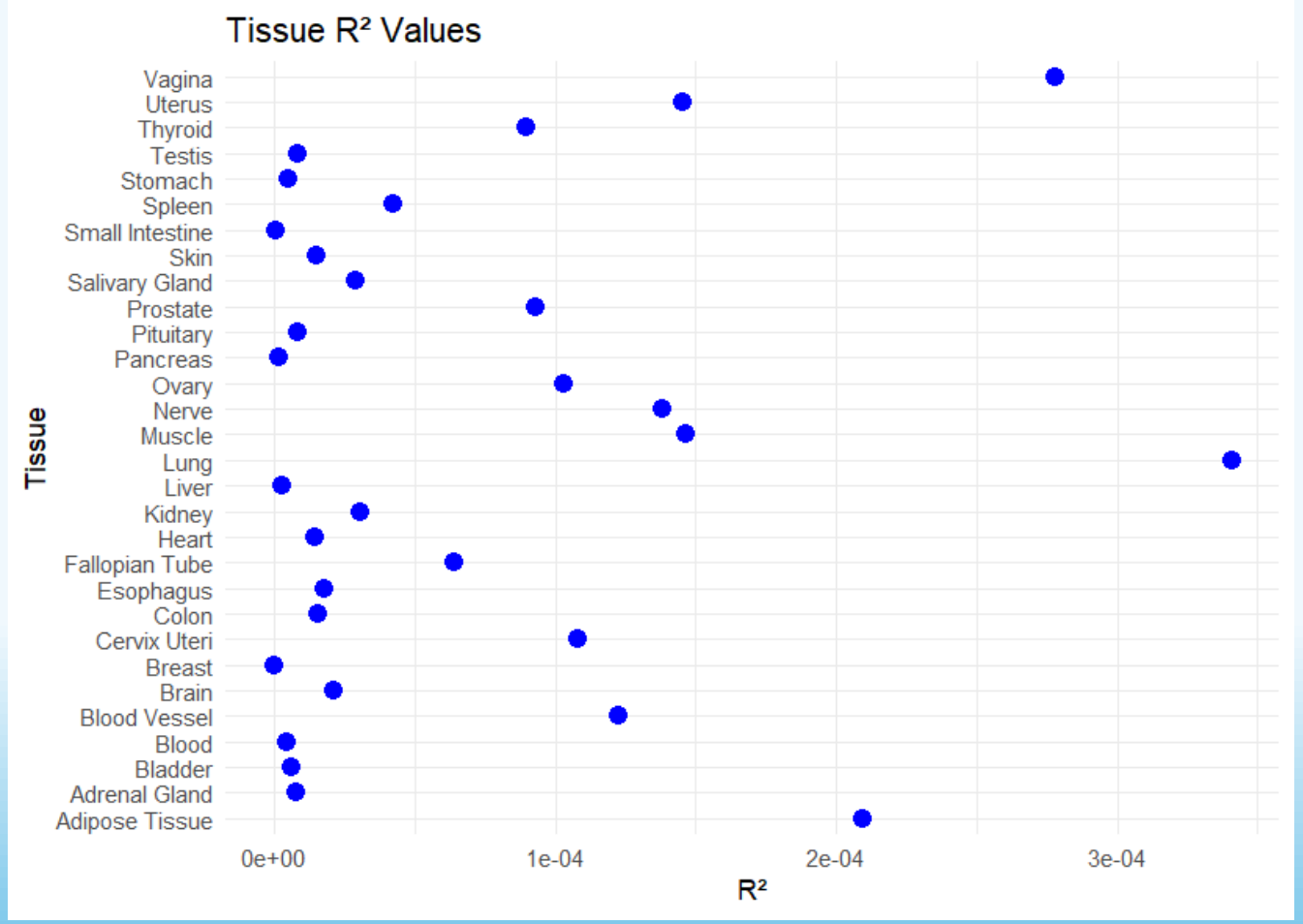
- There were multiple ways that we conducted our analysis, the previous results being focused around each individual tissue and how age correlates with it

- In this way we have all of the tissues grouped together under their overall type (i.e. Brain, Lung, etc.) and have different results for significant age correlations.

	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

Distribution of R^2 Values

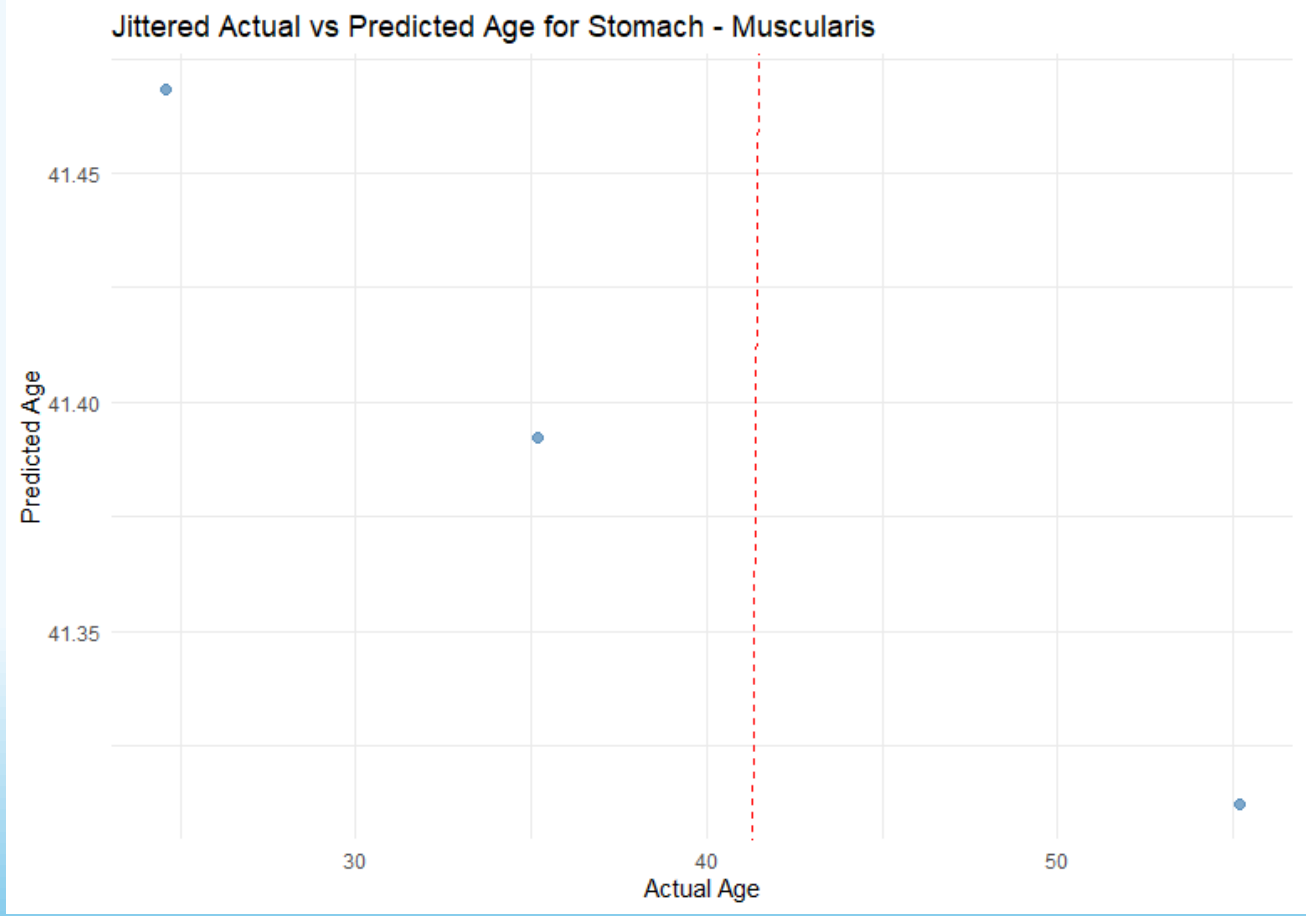
- However this proves to make an ineffective model fit as show by the insignificant R^2 values.



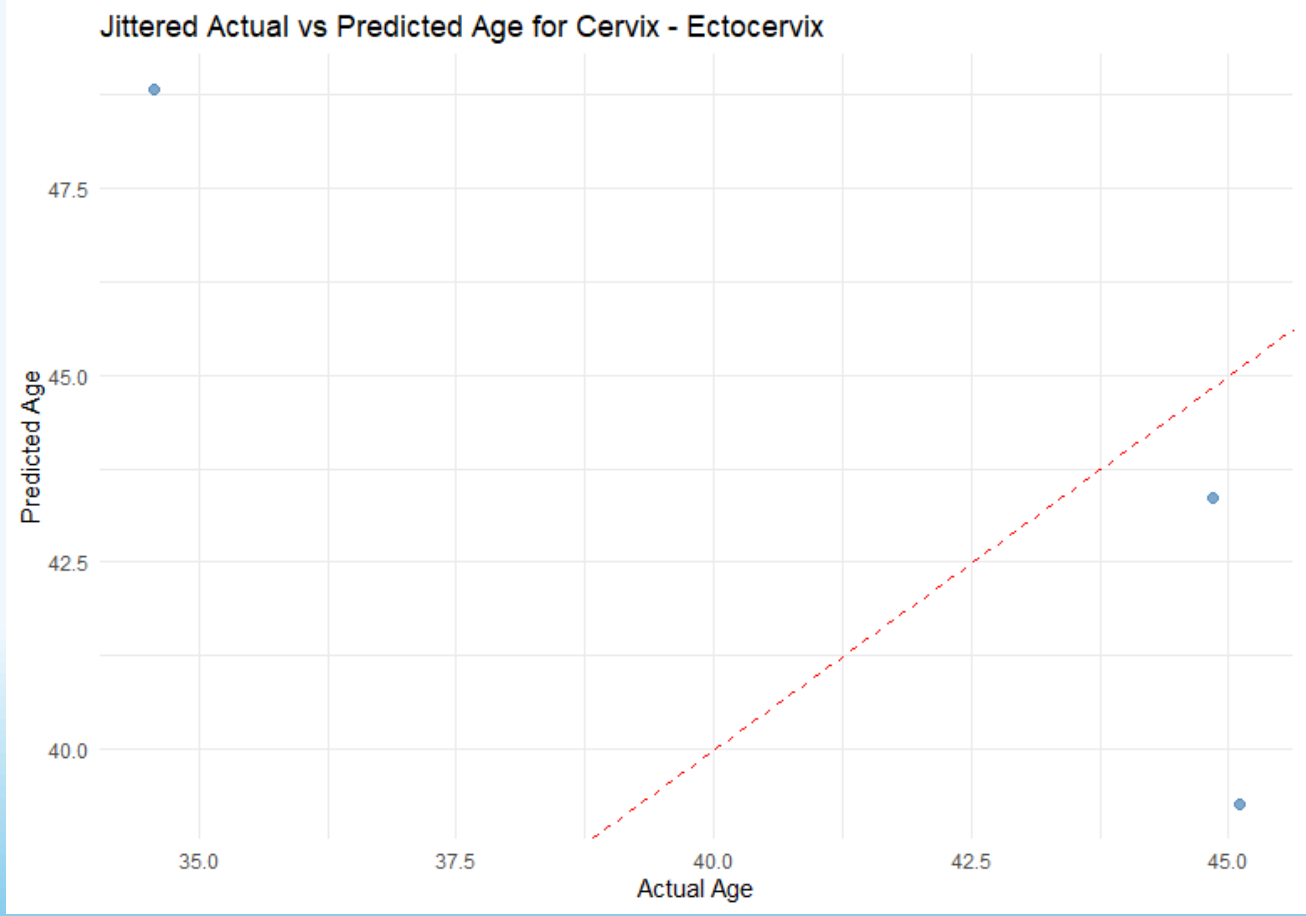
KNN Prediction

- Data split into training and testing sets, each randomly selected
- Number of samples available per tissue presents an issue
  - Some tissues had few total samples, so there were few candidates for training or testing data
  - i.e. Stomach- Muscularis had only 26 total samples, while muscle-skeletal had 818 samples
  - Distribution of samples across age groups may impact prediction power
- High susceptibility of gene expression to environmental impacts may impact accuracy of predictions

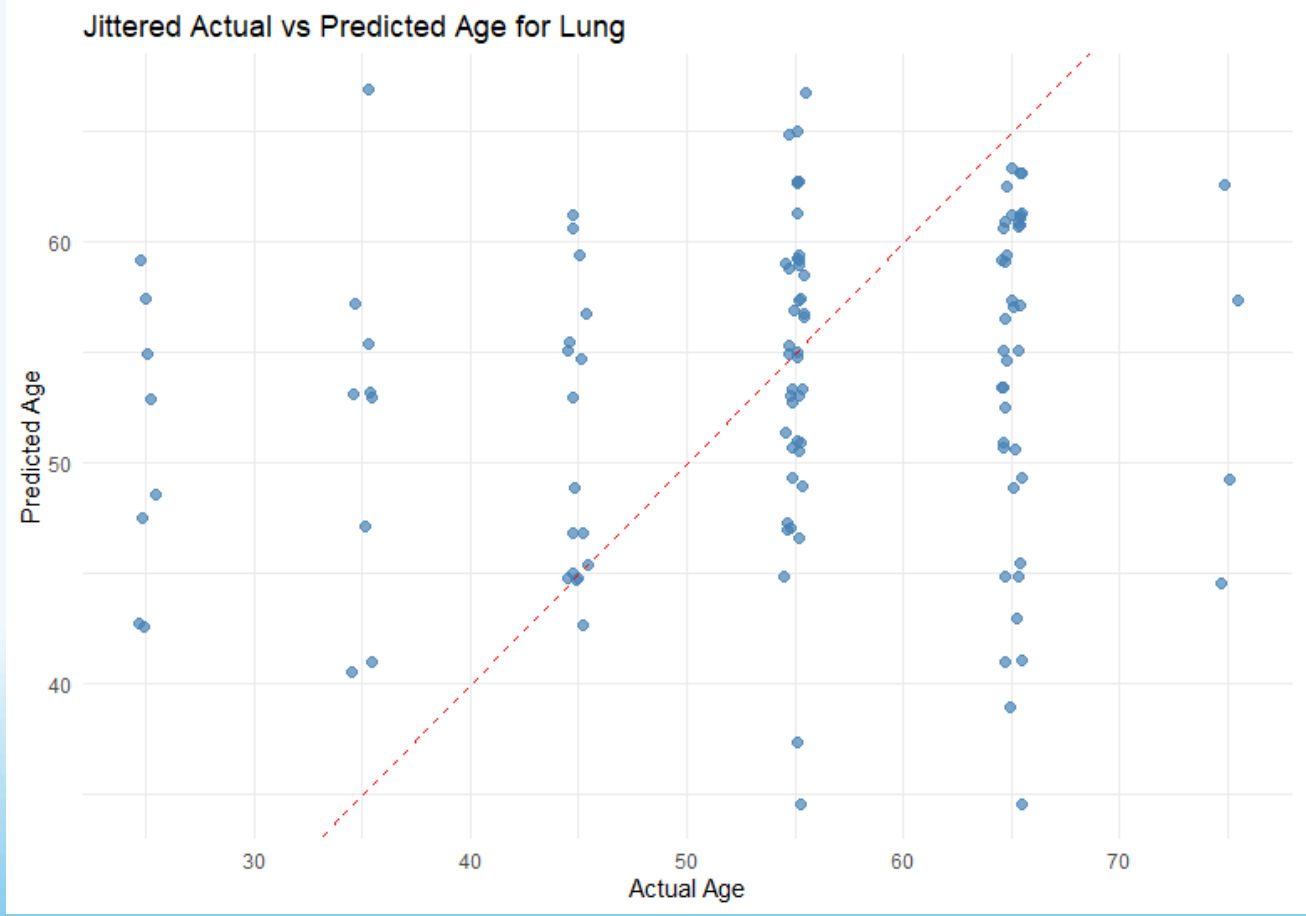
KNN Prediction: Stomach - Muscularis



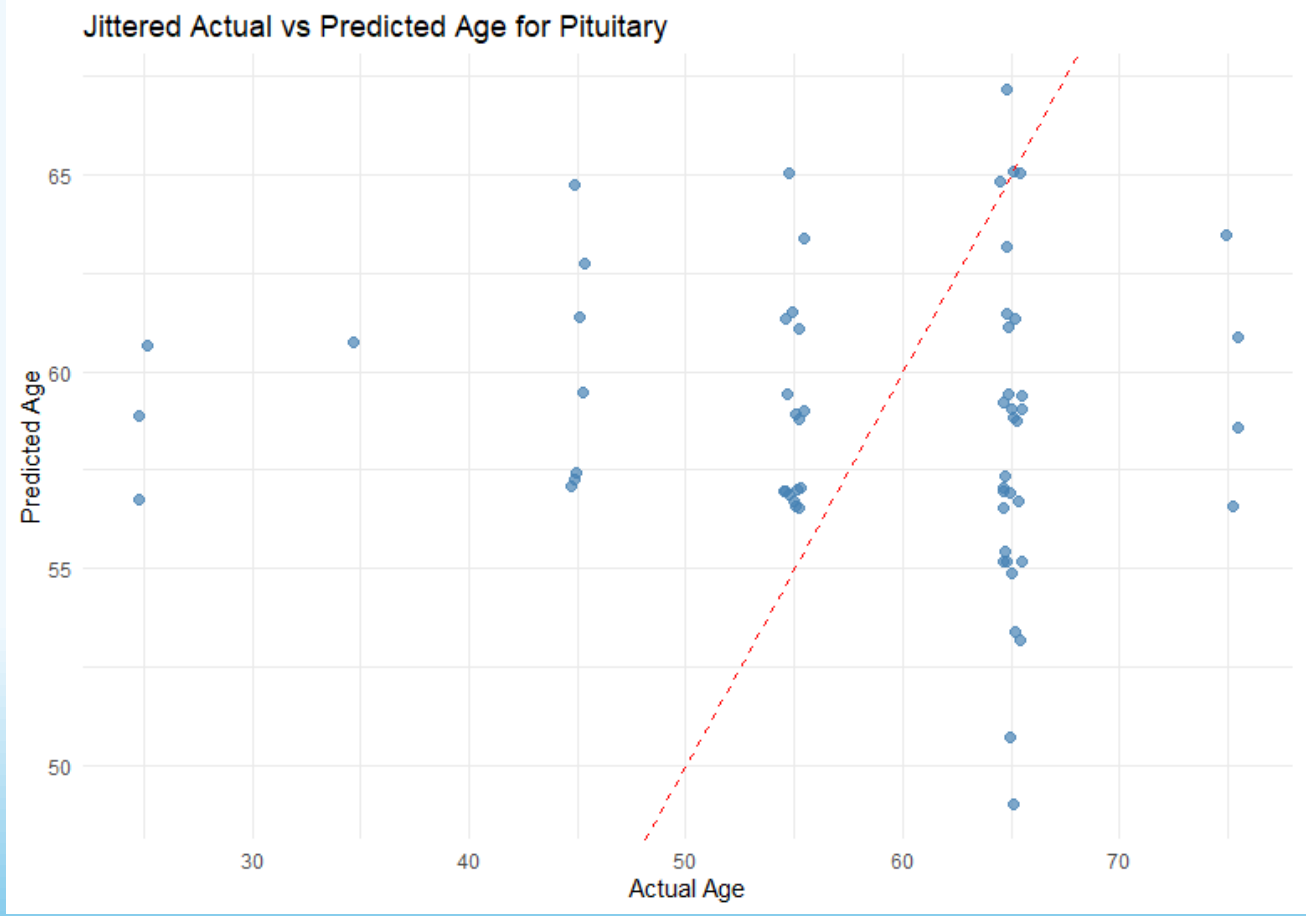
KNN Prediction: Cervix - Ectocervix



KNN Prediction: Lung



KNN Prediction: Pituitary



Conclusions

- While there are genes that do correlate with age, generally gene expression data is not a useful predictor for subject age.
- Moderate correlation statistics indicate that, while there is a moderate relationship, gene expression is not useful on its own.
- Human genetics are highly diverse among individuals. While one individual may exhibit “normal” expression levels for their age group, another individual in the same age group may have wildly different expression levels.
- Environmental factors can impact gene expression. The lack of control over exposure drastically reduces the feasibility of using gene expression as a predictor of subject age.

Sources

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