A Project Module Report for M. Sc. in Computational Biology

Pseudo-Bulk Conversion of scRNA-seq data, and Construction and Validation of Mouse Lung Transcriptomic Clock

Submitted by,

Santhosh Gojjam Kantharaju

Under the Guidance of,

Prof. Dr. Bjorn Schumacher

Department of Biology

University of Cologne, Germany

Contents

1 Abstract

Transcriptomic aging clocks estimate an organism's biological or chronological age based on high-dimensional profile data. Although this method has worked well for a variety of species and tissues, the possibilities of training transcriptomic clocks using scRNA-seq data have not yet been thoroughly investigated. In order to solve this, we converted scRNA-seq data into pseudo-bulk RNA-seq data using an aggregation method and investigated the impact of cell type composition on age prediction. Our research involved building two aging clocks. Using scRNA-seq data from mouse lungs that covered all cell types, the first clock examined the impact of aging on several cell types within the same organ. The second clock solely employed mouse lung epithelial cells, concentrating on a particular cell type.

We developed a transcriptomic clock using elastic-net regression techniques and utilized converted pseudobulk RNA-seq data to train the model. We also employed computational techniques such as feature selection and quantization in our analysis. Our model demonstrates a high performance in predicting biological age, achieving an accuracy of 97% (R2-Score) across all lung cell types. However, it performed poorly, with only 30% (R2-Score) accuracy, for a model focused exclusively on lung epithelial cells. Additionally, we validated our model using other datasets to assess the performance of age prediction. This work enables us to explore both global and cell-type-specific effects on aging.

Keywords: Single-cell RNA sequencing, Pseudo-bulk Aggregation, Aging Clocks, Quantilization

2 Introduction

2.1 Omics Data

2.1.1 Single-Cell RNA Sequencing Data

Single-cell RNA sequencing (scRNA-seq) has emerged as one of the most widely used genomic tools for analyzing transcriptomic heterogeneity and identifying rare cell types and states[12]. A key component of scRNA-seq analysis is the expression matrix, which represents the number of transcripts detected for each gene in each cell[2]. The elements of this matrix indicate the number of molecules assigned to a specific gene and cell[8]. scRNA-seq allows researchers to quantify the transcriptomes of thousands of individual cells simultaneously. These experiments often involve multiple subjects, cell lines, or biological replicates, enabling scientists to explore transcriptomic changes across different conditions[9].

2.1.2 Bulk RNA Sequencing Data

Bulk RNA sequencing (bulk RNA-seq) of heterogeneous mixtures represents average expression levels for each gene across the different cell types present in the mixture [6]. Bulk RNA-seq involves sequencing two types of libraries: messenger RNA (mRNA)-only libraries and whole transcriptome libraries, which include all RNA species except for ribosomal RNA (rRNA). This sequencing method is straightforward and cost-effective, primarily focusing on mRNA. One of the disadvantages of bulk RNA-seq is that the signals driving specific cell types can be blurred by an average gene expression profile derived from bulk RNA-seq. These signals are captured in each cells in scRNA-seq[12].

2.2 Pseudo-bulk Aggregation

RNA expression is a reflection of the distinct transcriptional regulation and epigenetic changes that occur at the RNA level in cells. Additionally, RNA is more susceptible to the influence of both micro and macro-environmental stimuli. Given the remarkable transcriptional diversity observed at the single-cell level, it is essential for cells to receive individualized treatment based on transcriptomic profiles. This need has spurred the development of various scRNA-seq technologies[12].

Using an aggregation approach, pseudo-bulk RNA-seq data is created by converting scRNA-seq data into bulk-like data that contains all of the cell-specific information. This is achieved by aggregating gene counts within each cell type and subject. For pseudo-bulk methods, aggregation of count values can be performed using two approaches: cumulative summation of raw count values (sum) or averaging single-cell normalized count values (mean). Junttila et al.[9] performed a comprehensive comparison of 18 methods, which include a naive single-cell method, pseudo-bulk method, and mixed models, for the identification of differential state changes between conditions from multisubject scRNA-seq data. They concluded pseudo-bulk methods performed generally well, and in pseudo-bulk method sum aggregation performed better than mean aggregation.

2.3 Age Prediction Clocks

Aging clocks are typically developed using supervised machine learning techniques. At the core of many modern modeling tools is linear regression, which provides a straightforward approach by representing the regression function as a linear combination of predictors. This linear structure makes the model parameters easy to interpret. The performance of a model can be evaluated using two key concepts: bias and variance. Bias, also known as mean squared error (MSE), refers to the squared error that occurs when the data used to train the model skews the output away from the expected result. On the other hand, variance measures the variability in the model's predictions.The performance of a model can be evaluated using two key concepts: bias and variance. Bias, also known as mean squared error (MSE), refers to the squared error that occurs when the data used to train the model skews the output away from the expected result. On the other hand, variance measures the variability in the model's predictions. [16]. Additionally, R2 scores measure the percentage of variance in the test data that each model can explain [15].

Several strategies can be used to improve the performance of the regression mode like, feature selection, regularization techniques, and cross validation.

- Feature selection is a crucial step in improving the performance of a model. Often, the features in a dataset are not suitable for processing by subsequent modules due to high dimensionality. The challenges associated with high-dimensional data, commonly referred to as the "curse of dimensionality" in statistical learning, indicate that the number of training samples must increase exponentially with the number of features to learn an accurate model. Since only a limited number of samples are usually available, there is an optimal number of feature dimensions [7].

- Regularization techniques are used to enhance model performance by introducing penalties that stabilize the model. Ridge linear regression applies an L2-norm penalty by adding the squared values of the coefficients to the loss function. In contrast, Lasso linear regression employs an L1-norm penalty, which incorporates the absolute values of the coefficients into the loss function. Elastic Net is a hybrid method that combines both regularization techniques and variable selection. Ridge regression is advantageous in reducing redundancy by addressing high multicollinearity issues, while Lasso regression is advantageous for feature selection. The L1 ratio determines the proportion of each regularization method used[1]. Hence in this project module, we aim to build different models using different regularization techniques, to find the best-performing model according to its MSE and R2 score.

- Hyperparameter tuning is essential for enhancing model performance, but it can lead to significant overfitting. Therefore, it's common practice to divide the available data into training and validation sets. The model that shows the best performance on the validation set is chosen as the final model. One effective method for this is K-fold cross-validation, which splits the data in a way that allows each sample to be used for both training and validation. In each iteration of the split, some examples are reserved for validation, while the rest are utilized for training[7].

Many age-prediction clocks have been published, initially starting with the patterns of DNA methylation levels of CpG sites to build the clock. With the advancement of RNA sequencing, transcriptomic data was used to identify transcriptomic aging biomarkers, creating an RNA expression signature for age classification. This was first introduced by Meyer et al 2021 [13], where the authors used bulk RNA-seq data of *C.elegans* with a binarization method that transforms the data into 0s and 1s based on medians to build a transcriptomic clock. A similar regression model was developed using microarray data of mouse lungs to determine whether cigarette smoke affected the premature aging of the mouse lungs [5]. To this day, there are a few clocks built using scRNA-seq data; one such clock was created using mouse scRNA-seq data, where an age classifier was developed for young and old mice [14].

Bulk tissue data fails to reveal the molecular aging processes in specific cell types, and it remains unclear how bulk aging signatures are influenced by variations in cell type composition [14]. To investigate this, we opted to use scRNA-seq data to develop a transcriptomic clock along with transformation techniques like quantilization, in which the training data was transformed into discrete bins based on quantiles. Our goal is to assess whether cell type composition impacts biological age predictions using the transcriptomic clock. For the clock's training, we utilized pseudo-bulk RNA-seq data. In this project module, we developed a model to predict biological age using pseudo-bulk transcriptomic data aggregated from scRNA-seq data. We applied various regression techniques and included a feature selection method. To minimize the mean squared error (MSE), we implemented nested cross-validation. Our results showed that the elastic-net regression model achieved the best performance. We validated the model's effectiveness by aggregating separately the available mouse aging lung scRNA-seq data Angelidis dataset[3], Kimmel dataset[10], and Tabula-Muris Facs and Droplet dataset[17].

3 Methods

3.1 Generation of the Training Data

Pseudo-bulk aggregation is a technique that converts scRNA-seq data into bulk-like RNA sequencing data. The pseudo-bulk methods combine count values from each sample and cell type (cluster) to create a dataset that can be analyzed using the same methods as bulk RNA-seq data. This maintains the same number of genes while it combines the number of cells and provides the number of samples in the gene expression matrix[9]. The below table1 provides information about the scRNA seq data obtained from UCSC Cell Browser [18] used for the training data.

UCSC Cell Browser Data	
Number of Cells	2,544,936 cells
Number of Genes	55416 genes
Age Cohort (No. of Mouse)	3 months(12), 6 months(8), 12 months(8), 16 months(11), 23 months(8)
Total Number of Mouse	47
Cell Types	Immune Cells, Stromal Cells, Epithelial Cells, Endothelial Cells, Adipose Cells

Table 1: ***Meta Information of Training data***, *Details of the mouse lung scRNA-seq dataset used for clock training.*

3.1.1 Pseudo-bulk of Whole Lung Data

Pseudo-bulk aggregation was performed on the scRNA-seq data from 2.5 million cells of mouse lung tissue obtained from UCSC Cell Browser[18]. Quality control was conducted on the scRNA-seq data to identify and remove poor-quality cells, including doublets, empty droplets, and dead cells.

A customized function was developed to aggregate the count values for pseudo-bulk analysis. This function performed a cumulative summation of the raw count values. The unique() function returns the unique values, meaning the function returns an array of non-repeated elements. This unique() function was employed to group all the mouse samples, and a loop was used to sum their raw count values.

In addition to the pseudo-bulk data, a metadata table was created using the gene names from adata.varnames() along with the IDs, sex, and age of the mice from adata.obs() in the scRNA-seq data.

We aggregated the raw count values using a cumulative summation approach. This means that all the raw count values from a single mouse were summed and recorded as one value. As a result, we obtained a matrix consisting of 47 mouse samples in the rows and 55416 genes in the columns.

3.1.2 Pseudo-bulk of Epithelial Cell Types

The epithelial cells were separated from the scRNA-seq data using the unique() function on the 'celltype' column in adata.obs. Following this, pseudo-bulk aggregation was performed, as explained in the previous section, along with the generation of a metadata table. The dimensions of this matrix were the same for both the whole lung dataset and the epithelial cell dataset.

3.2 Construction of Age Prediction Clock

3.2.1 Feature Selection

Selecting the right features is essential for enhancing model performance. Certain features might be inappropriate due to high dimensionality, a challenge often referred to as the "Curse of Dimensionality" in statistical learning [7].

To enhance performance, feature selection was conducted on the pseudo-bulk data to eliminate the least expressed and least variable genes.

First, to remove the least expressed genes, we eliminated those with a value of zero across all samples. Then, we removed genes with a mean expression level of less than 0.0001 raw reads counts.

Next, we focused on eliminating the least variable genes. To do this, we log-transformed the data and calculated the coefficient of variance for the log-transformed data. Only the top 25% of the most variable genes were retained, while the remaining genes were discarded.

3.2.2 Quantilization

To optimize the model's performance, a quantization step was performed. An in-house function was created to assign samples to quantile bins based on the input, in this case, 10 bins. The data was divided into 10 quantiles based on their row-wise values (mouse sample). As a result, the final output displays quantile bins instead of the original raw count values.

3.2.3 Regression Model of Whole Mouse Lung Clock and its Validation

The clock model was built using a nested cross-validation (CV) technique as shown in figure[1], where CV was performed twice, making it an efficient method for hyperparameter tuning. Both the outer and inner CV were conducted using 5 folds. In the inner CV, the model is trained, and only the best models, identified by their mean squared error (MSE) values, are selected based on the best alpha parameters. These selected alphas then undergo the outer CV, and the final model is chosen based on the lowest MSE and the highest R2 score.

Linear regression models were trained using training and testing sets obtained through K-Fold cross-validation. Four segments of the dataset were utilized for training, while one segment was reserved for testing. This process was repeated in a loop five times to assess all possible combinations of features. The MSE and R2 score for each iteration were calculated and recorded.

Additionally, linear models with L1 regularization, specifically Ridge and Lasso (from

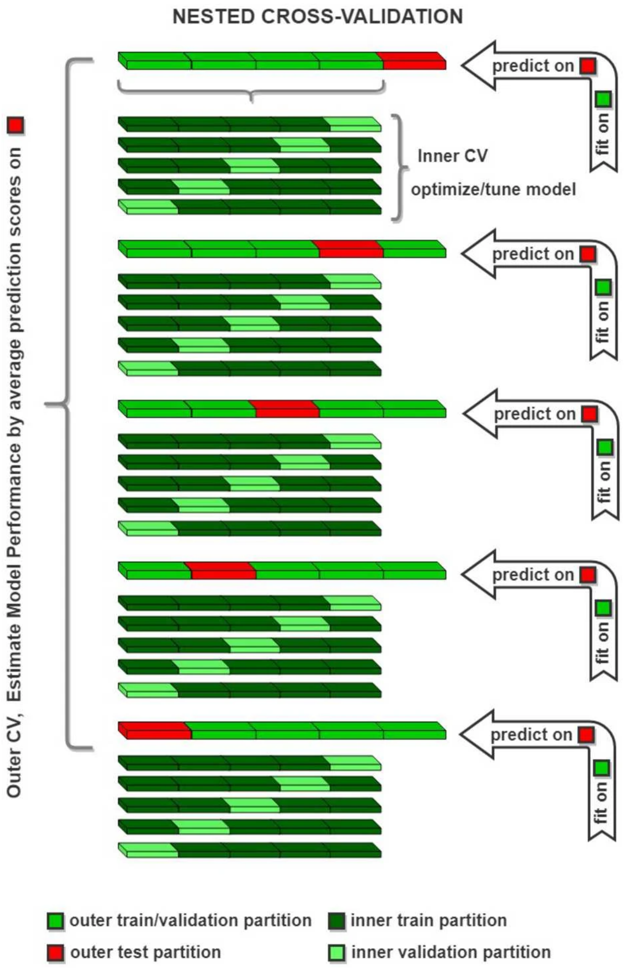


Figure 1: ***Nested Cross Validation representation***, *[11] which was used to build the Age Prediction Clock*

`sklearn.linear_model`), were trained using alpha parameters that ranged from 0 to 1 in increments of 0.1. Elastic Net (also from `sklearn.linear_model`), which combines both L1 and L2 regularization, was trained using the same alpha values, along with an L1 ratio that varied from 0.0 to 1.0 in intervals of 0.2. Hyperparameter tuning was conducted to identify optimal values for the alpha and L1 ratio, aiming to enhance the regression model's performance. The MSE and R2 scores for all models were calculated and documented.

3.2.4 Regression Model of Epithelial Mouse Lung Clock and its Validation

To construct the epithelial mouse lung clock, pseudo-bulk data of epithelial mouse lung cells was used. Different regression models were trained using the parameters as explained in the previous section, but with one alteration. Instead of using nested-CV, 5-fold cross-validation was used here. The MSE and R2 scores for all the models were calculated and recorded.

3.3 Validation of Age Prediction Clock

3.3.1 Validation of Whole Mouse Lung Clock

The coefficients and intercept values for the clock were determined. To validate the results, we selected common features (genes) that were present in both datasets. The validation was conducted using the formula $y = mx + c$, where y represents the predicted age for the mouse, m is the vector of coefficient of the genes derived from the clock, x is a matrix which contains the raw count values from the validation dataset, and c is the intercept.

We obtained various validation datasets from Angelidis Dataset[3], Kimmel Dataset[10], and Tabula-Muris Facs & Tabula-Muris Droplet[17], and performed pseudo-bulk aggregation using the same method as above. This approach was utilized to validate the clock, and a graph comparing chronological age (months) to predicted age (months) was plotted.

3.3.2 Validation of Epithelial Mouse Lung Clock

The coefficients and intercept from the best model were used for validation, which was also performed like the prior explanation. Epithelial cells from the datasets Angelidis Dataset[3], Kimmel Dataset[10], and Tabula-Muris Facs & Tabula-Muris Droplet[17] were extracted for pseudo-bulk aggregation. This pseudo-bulk data was then utilized for validation.

3.4 Data Availability

The scRNA-seq data used in this module was obtained from the following papers and databases listed in the below table 2.

Dataset	Total Number of Cells	Total Number of Mouse	Reference (doi)
UCSC Cell Browser	2,544,936 cells	47 Mice	10.1126/science.adn3949
Angelidis	14,170 cells	15 Mice	10.1038/s41467-019-08831-9
Kimmel	30,255 cells	13 Mice	10.1101/gr.253880.119
Tabula-Muris FACS	5,218 cells	14 Mice	10.1038/s41586-020-2496-1
Tabula-Muris Droplet	24,540 cells	16 Mice	10.1038/s41586-020-2496-1

Table 2: ***Metadata of Datasets Used for Construction and Validation****, scRNA-seq data from mouse lung used in this study include the total number of mice, the total number of cells, and the reference for the publication where the dataset was originally obtained used.*

4 Results

4.1 Feature Selection

Feature selection is conducted to reduce dimensionality and redundancy while improving model performance. To achieve this, we remove the least expressed and least variable genes. Eliminating low-expressing genes and least variable genes is crucial as they introduce noise into the data, allowing us to focus on the highly expressed genes responsible for age prediction. In contrast, highly variable genes tend to be more biologically informative.

We removed the least expressed genes whose mean expression across all mouse individuals was less than 0.0001. This action reduced the number of genes from 55416 to 26891. For the least variable genes, we first performed a log transformation to stabilize the variance for highly expressed genes, reduce the influence of outliers, and normalize the differences between low and high expression levels. After the log transformation, we removed the least variable genes based on their coefficient of variation, retaining only the top 25% of the most variable genes. This process further reduced the number of genes from 26891 to 20168. The resulting 20168 genes are highly variable and highly expressed, making them more biologically informative for age prediction. A comprehensive performance result is shown in figure [2].

4.2 Regression Model of Whole Mouse Lung Clock and its Validation

A regression model was trained employing optimal features. A scatterplot is presented in figure [3], where the x-axis represents the chronological age (months) and the y-axis displays the predicted age (months) according to the model. Each point on the scatterplot corresponds to a mouse sample, while the red line depicts the regression trend. Various models were trained with different regularization parameters detailed results are provided in table[3], and it was determined that the Elastic-Net Regression model, with an alpha value of 0.1, and l1 ratio of 0.6 achieved the best prediction performance. This model resulted in

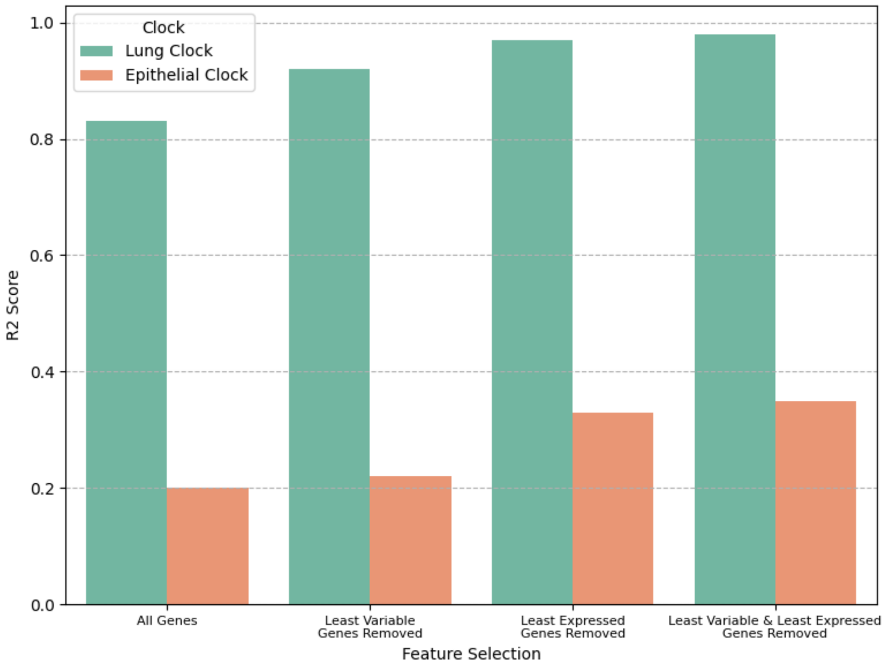


Figure 2: ***Feature Selection Results***, *Performance of the Clocks with different feature selection techniques, plotted against R2 score*

an R2 score of 0.98 and a mean squared error (MSE) of 0.96 months.

Model	MSE (months)	R2 Score
Linear Regression	5.09	0.91
Ridge Regression ($\alpha = 0.0$)	5.09	0.91
Lasso Regression ($\alpha = 0.6$)	1.30	0.97
Elastic-Net Regression ($\alpha = 0.1, l1 - ratio = 0.6$)	0.96	0.98

Table 3: ***Performance of Whole Mouse Lung Clock***, *Results of the mouse lung clock for different regression models with MSE and R2 score.*

Coefficients for the genes and intercepts were obtained and used to validate the clock. The scatter plot compares chronological ages (months) with the clock's predicted ages (months).

The plots in figure[4] indicate that the clock's performance on unseen data is non-

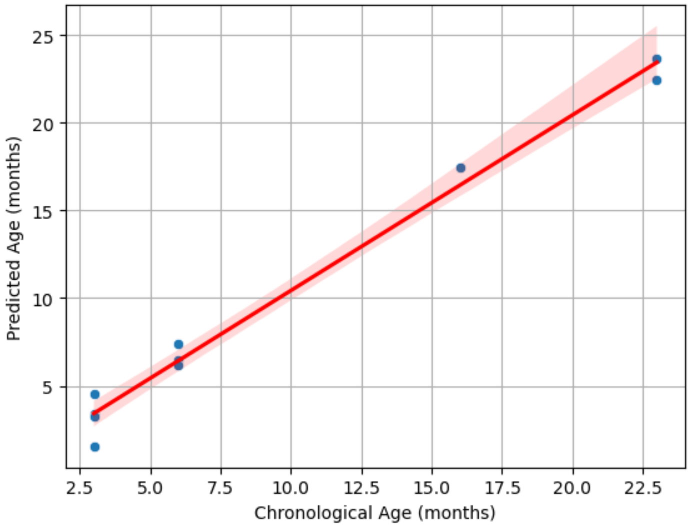


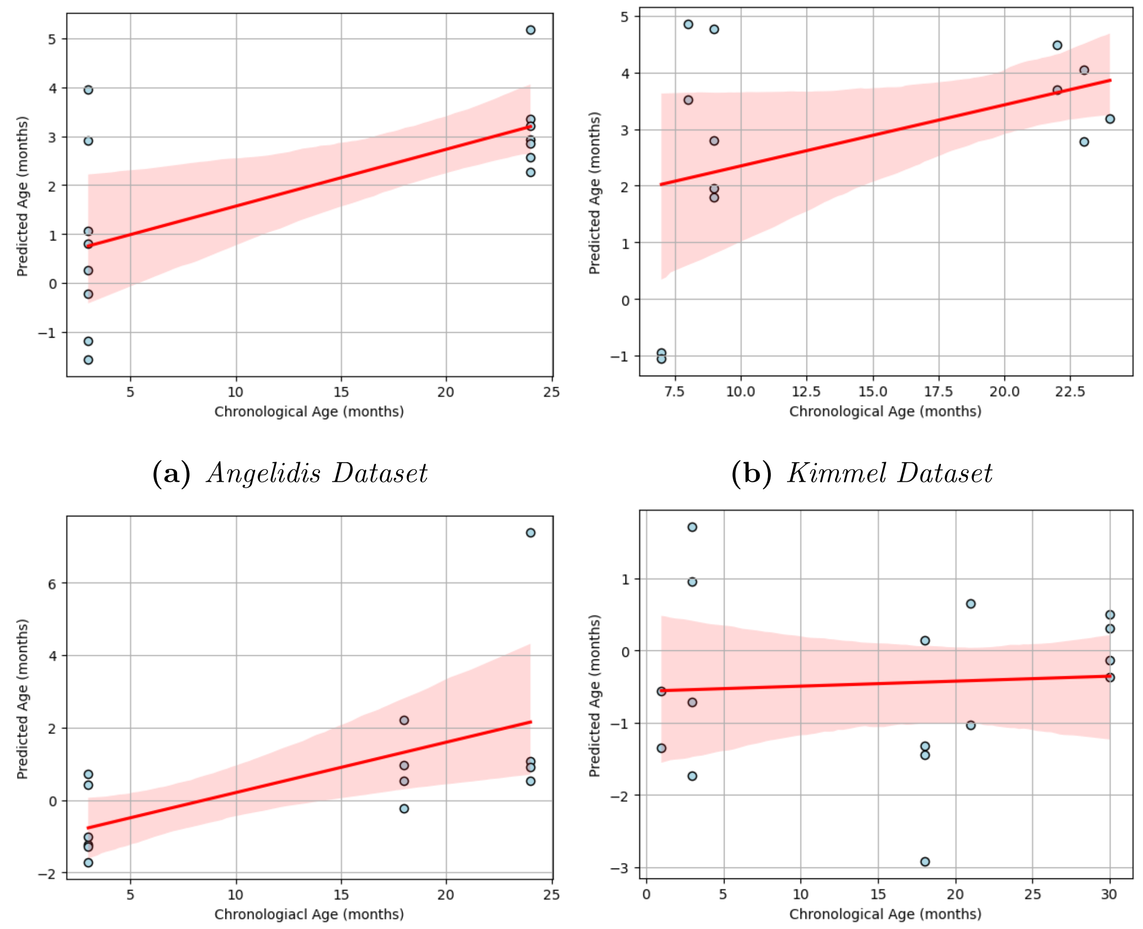
Figure 3: ***Performance of Whole Mouse Lung Clock****, Scatterplot of Predicted Age (months) vs Chronological Age (months), with the red line representing the regression line, blue dots representing the data points and red shaded region representing 95% confidence level. The model was build using Elastic-Net regression, alpha=0.1, and l1 ratio=0.6. The prediction performance of the model was R2 score=0.98*

satisfactory, leading to an overall poor validation. The performance of the clock is poor as it is predicting negative ages and there are huge differences between chronological age (months) and predicted age (months).

4.3 Regression Model of Epithelial Mouse Lung Clock and its Validation

A regression model was developed using optimal features, detailed results are provided in table[4]. A scatterplot figure[5] depicts the performance of the Elastic-Net Regression model, with an alpha value of 0.1 and l1 ratio of 0.0, achieving the best prediction performance. This model obtained an R2 score of 0.35 and a mean squared error (MSE) of 30.06 months.

Coefficients for the genes and the intercepts were obtained and used to validate the clock. The scatter plot illustrates the chronological ages (months) compared to the predicted



(a) *Angelidis Dataset*

(b) *Kimmel Dataset*

(c) *Tabula Muris Facs Dataset*

(d) *Tabula Muris Droplet Dataset*

Figure 4: ***Validation Results of Whole Mouse Lung Clock***, *Scatterplot of Predicted Age (months) vs Chronological Age (months), with the red line representing the regression line and blue dots representing the data points. The Mouse Lung Clock validation for the different datasets.*

ages (months) from the clock.

However, the plots in figure[6] below indicate that the clock's performance with the unseen data is insufficient and overall poor. In the (b) Kimmel Dataset and the (c) Tabula Muris Facs Dataset, there is a negative correlation, while in other datasets, the difference between predicted age (months) and chronological age (months) is significant.

Model	MSE (months)	R2 Score
Linear Regression	34.68	0.23
Ridge Regression ($\alpha = 0.0$)	34.68	0.23
Lasso Regression ($\alpha = 0.1$)	48.69	-0.06
Elastic-Net Regression ($\alpha = 0.1, l1 - ratio = 0.0$)	30.06	0.35

Table 4: ***Performace of Epithelial Mouse Lung Clock***, *Results of the mouse epithelial lung clock for different regression models with MSE and R2 score.*

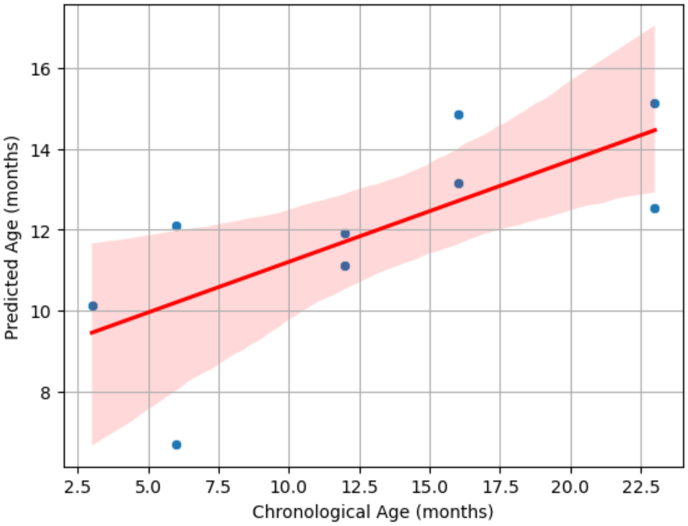
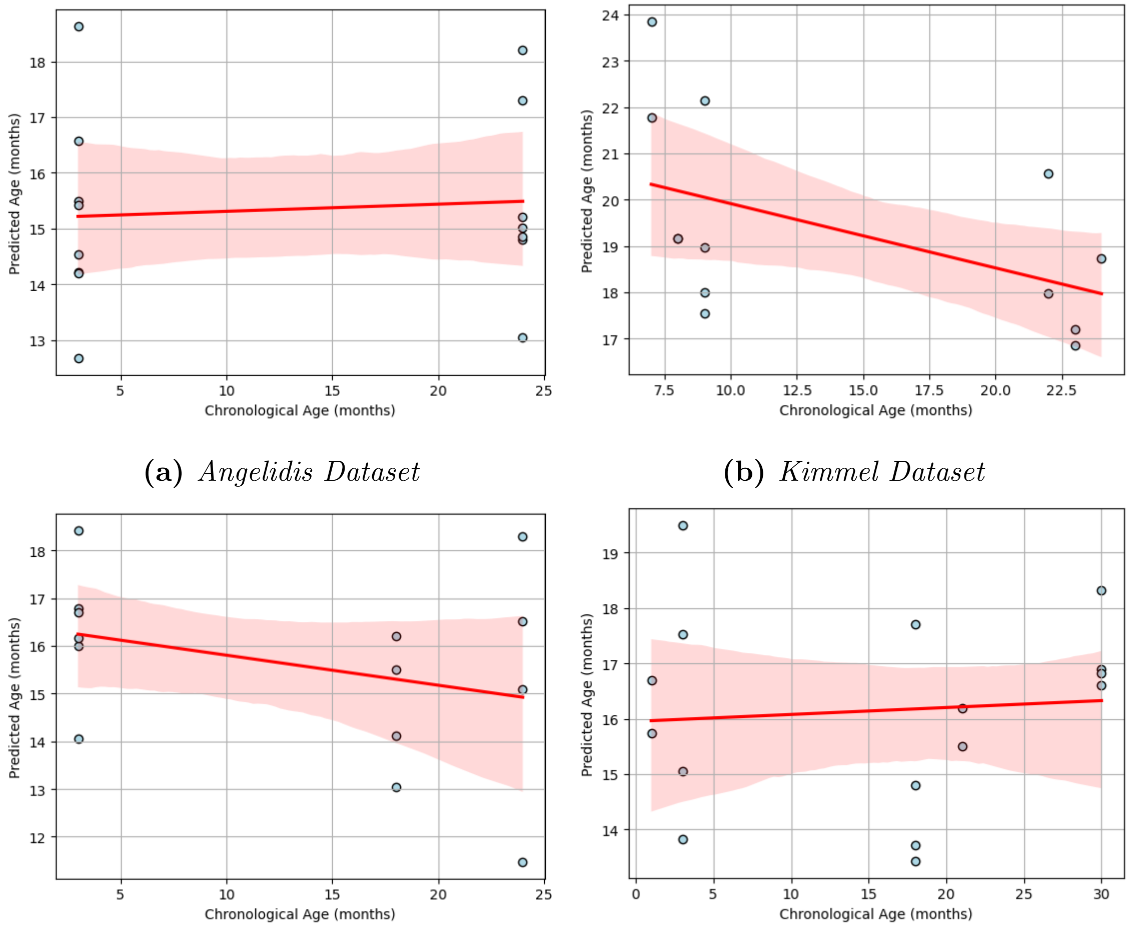


Figure 5: ***Result of Epithelial Mouse Lung Clock***, *Scatterplot of Predicted Age vs Chronological Age, with a red regression line, blue data points, and a red shaded 95% confidence region. The Elastic-Net regression model (alpha=0.1, l1 ratio=0.0) achieved an R2 score=0.35*



(a) *Angelidis Dataset*

(b) *Kimmel Dataset*

(c) *Tabula Muris Facs Dataset*

(d) *Tabula Muris Droplet Dataset*

Figure 6: ***Validation Results of Epithelial Mouse Lung Clock****, Scatterplot of Predicted Age (months) vs Chronological Age (months), with the red line representing the regression line and blue dots representing the data points. The Mouse Lung Epithelial Clock validation for the different datasets.*

5 Discussion

Aging is a complex process that affects multiple biological systems. We need specific evidence to confirm whether the transcriptomic changes observed in biological clocks are driving the aging process. Furthermore, it remains uncertain how differences between predicted age and chronological age contribute to biological age or prediction error. Clocks that are developed using small sample sizes may be influenced by confounding factors related to cell composition. Additionally, the reliability of these clocks across different environments, populations, and tissues is still not well understood[4].

Currently, no single model can accurately predict the biological age of various organisms across different strains, treatments, and conditions [13]. Additionally, no model has been trained using scRNA-seq data [14]. Choukrallah et al. [5] developed a regression model to predict age using microarray data from mouse lungs, applying LASSO regression, which resulted in an average prediction error of 0.83 months. Our goal is to utilize transcriptomic data to identify biomarkers of transcriptomic aging. A drawback of their study is that they used mouse samples from 2 months to 10 months, which means the model is trained only on young mice and cannot adequately explain the aging signatures in older mice. Meyer et al. [13] successfully built an age prediction model using bulk RNA-seq data from *C. elegans*, employing a transformation technique called binarization. However, bulk RNA-seq does not adequately capture the molecular aging processes in specific cell types, nor the aging signatures that are influenced by variations in cell type composition. To address this, Neumann et al. [14] used scRNA-seq data to develop an age classifier, implementing elastic-net and tree-based machine learning techniques such as random forests and XGBoost. One major limitation they encountered was the small number of individual organisms in their dataset. Their model was trained and validated on only one dataset, although they attempted to mitigate this limitation with a sophisticated training and testing setup. Ultimately, they were only able to differentiate between young and old mice and did not train regression models to predict exact ages.

To address the limitations and challenges identified by Choukrallah et al. [5], and Neumann et al. [14], We utilized a larger scRNA-seq dataset with a distributed age cohort that includes both young and old mice, ranging from 3 to 23 months, and performed pseudo-bulk aggregation to convert it into bulk-like data. We then trained an elastic-net regression model to predict age. Additionally, we employed quantization, where we utilize multiple quantiles instead of the two quantiles used in the binarization method, as described by Meyer et al., [13]. This technique transforms data into more than two discrete bins based on quantiles. By reducing noise while retaining important information, this approach may enhance the identification process.

We conducted various types of feature selection in combination with quantilization technique to enhance the prediction of biological age. Feature selection is important as multiple features (in our case, genes) are not suitable due to high dimensionality and redundancy. The problem of redundancy is also called collinearity in statistics. When two or more features are collinear, the covariance matrix of the dataset becomes singular and, therefore, irreversible, which leads to various problems [7]. In our case, genes found in one dataset may not be present in the validation set, or gene expression might be abundant in one dataset, but the gene expression for the same gene might not be abundant in other datasets. Hence, it is important to screen for the genes that do not pose problems in redundancy, which may be important for the better performance of the model. Our findings indicated that the best performance was achieved when we employed feature selection focused on highly expressed genes, as well as dispersion-based feature selection, which involved removing genes with low expression and low variability.

The validation of the clock highlights the challenge of accurately predicting unseen data. This difficulty may stem from several factors, including the significant impact of batch effects, which is a major issue when working with RNA-seq data, as well as the limited amount of training data. Although we have 2.5 million cells in our training set of scRNA-seq data, this data was obtained from only 47 mouse individuals. Consequently, the dimensionality of pseudo-bulk RNA-seq is drastically reduced to 47×55416, which is insufficient for training a machine learning model as it can lead to overfitting, despite the use of nested cross-validation.

A significant limitation of our study is that we only included five specific age cohorts for training the model: 3, 6, 12, 16, and 23 months. As a result, it is difficult to accurately predict ages for other cohorts or for continuous age ranges. This issue is evident in the validation results from the Kimmel dataset, where the model struggled to make predictions for previously unseen age cohorts, specifically those of 7, 8, and 9 months.

When the clock was trained solely on epithelial cells, its performance deteriorated further. This decline can suggest that the age prediction is not significantly contributed only by epithelial cells, but rather all types of cell types included. Therefore, we can conclude that using specific cell compartments for predicting the clock is not a favorable choice, as it decreases variability and allows other confounding factors to negatively affect performance. It may be preferable to use different strategies to account for variations in cell-type composition when generating pseudo-bulk samples. For instance, weighted aggregation involves assigning specific weights to each cell type before performing the pseudo-bulk aggregation. Alternatively, using only raw counts for pseudo-bulking can result in expression values that are dominated by the most abundant cell types. Additionally, old mice may exhibit different cell type abundances compared to young mice. Therefore, it is advisable to normalize within each cell type first, followed by aggregating the normalized raw counts to create the pseudobulk. This approach will ensure that the final pseudobulk reflects a normalized proportion of cell types, regardless of age.

One possible hypothesis is that combining two or more datasets with varying age points could enhance model performance and offer a continuous age spectrum for predicting biological age. This strategy would expand the size of the training set and introduce greater variability, which can help the model learn more effectively and reduce the risk of overfitting.

Our work concludes that scRNA-seq offers more advantages compared to bulk RNA-seq. scRNA-seq accounts for cell type composition and helps capture aging signatures specific to different cell types. However, scRNA-seq also presents several disadvantages. It involves complex datasets and requires more computational power to analyze them. A significant issue is gene dropout, where transcripts may go undetected in one dataset but not in another, potentially introducing bias during age prediction.

To address these disadvantages, pseudo-bulk RNA-seq emerges as a better option. Pseudo-bulk combines the advantages of bulk RNA-seq, being computationally easier to manage and producing greater reproducibility across different datasets, while also preserving the cell-type-specific information fundamental in scRNA-seq.

6 Bibliography

References

[1] Abdullah S. Al-Jawarneh, Mohd Tahir Ismail, and Ahmad M. Awajan. Elastic net regression and empirical mode decomposition for enhancing the accuracy of the model selection. *International Journal of Mathematical, Engineering and Management Sciences*, 6:564–583, 4 2021.

[2] Tallulah S. Andrews, Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. Tutorial: guidelines for the computational analysis of single-cell rna sequencing data, 1 2021.

[3] Ilias Angelidis, Lukas M. Simon, Isis E. Fernandez, Maximilian Strunz, Christoph H. Mayr, Flavia R. Greiffo, George Tsitsiridis, Meshal Ansari, Elisabeth Graf, Tim Matthias Strom, Monica Nagendran, Tushar Desai, Oliver Eickelberg, Matthias Mann, Fabian J. Theis, and Herbert B. Schiller. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10, 12 2019.

[4] Christopher G. Bell, Robert Lowe, Peter D. Adams, Andrea A. Baccarelli, Stephan Beck, Jordana T. Bell, Brock C. Christensen, Vadim N. Gladyshev, Bastiaan T. Heijmans, Steve Horvath, Trey Ideker, Jean Pierre J. Issa, Karl T. Kelsey, Riccardo E. Marioni, Wolf Reik, Caroline L. Relton, Leonard C. Schalkwyk, Andrew E. Teschendorff, Wolfgang Wagner, Kang Zhang, and Vardhman K. Rakyan. Dna methylation aging clocks: Challenges and recommendations, 11 2019.

[5] Mohamed Amin Choukrallah, Julia Hoeng, Manuel C. Peitsch, and Florian Martin. Lung transcriptomic clock predicts premature aging in cigarette smoke-exposed mice. *BMC Genomics*, 21, 4 2020.

[6] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E. Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications*, 11, 12 2020.

[7] Ricardo Gutierrez-Osuna. Pattern analysis for machine olfaction: A review, 2002.

[8] Christoph Hafemeister and Florian Halbritter. Single-cell rna-seq differential expression tests within a sample should use pseudo-bulk data of pseudo-replicates, 3 2023.

[9] Sini Junttila, Johannes Smolander, and Laura L. Elo. Benchmarking methods for detecting differential states between conditions from multi-subject single-cell rna-seq data. *Briefings in Bioinformatics*, 23, 9 2022.

[10] Jacob C. Kimmel, Lolita Penland, Nimrod D. Rubinstein, David G. Hendrickson, David R. Kelley, and Adam Z. Rosenthal. Murine single-cell rna-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Research*, 29:2088–2103, 2019.

[11] E. Lavasa, G. Giannopoulos, A. Papaioannou, A. Anastasiadis, I. A. Daglis, A. Aran, D. Pacheco, and B. Sanahuja. Assessing the predictability of solar energetic particles with the use of machine learning techniques. *Solar Physics*, 296, 7 2021.

[12] Xinmin Li and Cun Yu Wang. From bulk, single-cell to spatial rna sequencing, 12 2021.

[13] David H. Meyer and Björn Schumacher. Bit age: A transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell*, 20, 3 2021.

[14] Janis Frederick Neumann, Ana Carolina Leote, Meike Liersch, and Andreas Beyer. Predicting murine age across tissues and cell types using single cell transcriptome data, 10 2022.

[15] Srikari Rallabandi and Sourav Yadav. Comparing ridge and logistic regression with

neural networks. pages 1065–1073. Institute of Electrical and Electronics Engineers Inc., 2023.

[16] Xiaogang Su, Xin Yan, and Chih Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4:275–294, 5 2012.

[17] The Tabula and Muris Consortium. A single cell transcriptomic atlas characterizes aging tissues in the mouse hhs public access. *Nature*, 583:590–595, 2020.

[18] Zehao Zhang, Chloe Schaefer, Weirong Jiang, Ziyu Lu, Jasper Lee, Andras Sziraki, Abdulraouf Abdulraouf, Brittney Wick, Maximilian Haeussler, Zhuoyan Li, Gesmira Molla, Rahul Satija, Wei Zhou, and Junyue Cao. A panoramic view of cell population dynamics in mammalian aging. *Science*, 387, 1 2025.