Original Manuscript

Title Gene Age Gap Estimate (GAGE) for major depressive disorder: a penalized biological age model using gene expression

[1]Yijie (Jamie) Li, [2]Rayus Kuplicki, [3]Bart N. Ford, [1]Elizabeth Kresock, [2]Leandre Figueroa-Hall, [2,4]Jonathan Savitz, [1,5,*]Brett A. McKinney

[1]Tandy School of Computer Science, The University of Tulsa, Tulsa, OK, USA

[2]Laureate Institute for Brain Research, Tulsa OK, USA

[3]Department of Pharmacology and Physiology, Oklahoma State University Center for Health Sciences, Tulsa, OK, USA.

[4]Oxley College of Health and Natural Sciences, The University of Tulsa, Tulsa OK, USA.

[5]Department of Mathematics, The University of Tulsa, Tulsa, OK, USA

[*]corresponding, brett-mckinney@utulsa.edu

Abstract

Recent associations between Major Depressive Disorder (MDD) and measures of premature aging suggest accelerated biological aging as a potential biomarker for MDD susceptibility or MDD as a risk factor for age-related diseases. Statistical and machine learning regression models of biological age have been trained on various sources of high dimensional data to predict chronological age. Residuals or “gaps” between the predicted biological age and chronological age have been used for statistical inference, such as testing whether an increased age gap is associated with a given disease state. Recently, a gene expression-based model of biological age showed a higher age gap for individuals with MDD compared to healthy controls (HC). In the current study, we propose a machine learning approach that simplifies gene selection by using a least absolute shrinkage and selection operator (LASSO) penalty to construct an expression-based Gene Age Gap Estimate (GAGE) model. We construct the LASSO-GAGE (L-GAGE) model in an RNA-Seq study of 78 unmedicated individuals with MDD and 79 HC and then test for accelerated biological aging in MDD. When testing L-GAGE association with MDD, we account for factors such as sex and chronological age to mitigate regression to the mean effects. The L-GAGE shows higher biological aging in MDD subjects than HC, but the elevation is not statistically significant. However, when we dichotomize chronological age, the interaction between MDD status and age is significant in L-GAGE model. This effect remains statistically significant even after adjusting for chronological age and sex. We find cytomegalovirus (CMV) serostatus is associated with elevated L-GAGE. We also investigate feature selection methods Random Forest and nearest neighbor projected distance regression (NPDR) to characterize age related genes, and we find functional enrichment of infectious disease and SARS-COV pathways.

1. Introduction

Major depressive disorder (MDD) has been hypothesized to show characteristics of premature aging [1]. Biological aging can be measured in multiple dimensions such as telomere length, immunosenescence, brain volume, and gene expression. These measures of biological aging are correlated with chronological age, but environmental and genetic factors can increase or decrease an individual's biological age relative to their chronological age and influence their risk for age related diseases. For example, MDD has been associated with markers of cellular and immune aging including shortened leukocyte telomere length [2, 3], elevated indicators of oxidative stress[4], and elevated circulating inflammatory cytokines [5]. Epigenetic clocks predicting biological age based on the accumulation of methylated CpG sites have found higher biological age in MDD subjects compared with healthy controls [6]. Brain age models constructed from T1-weighted magnetic resonance image (MRI) data from 2,188 healthy controls predicted a gap of +1.08 years (SE 0.22) between predicted and chronological age across 2,675 depressed subjects [7].

A recent RNA-Seq MDD study from Cole at el. found that gene expression based biological aging was elevated in MDD subjects compared to HC [8]. The PBMC samples included four groups: 44 healthy controls, 94 MDD treatment-resistant, 47 MDD treatment-responsive and 46 MDD untreated [8]. They selected age genes iteratively by varying the P-value threshold for the t-test between upper and lower chronological age quartiles. For a given iteration, a biological age was computed for each subject based on the signed z-score of the age-related genes, and the P-value threshold was chosen to optimize the correlation between biological and chronological age of the subjects (Spearman Correlation Coefficient (SCC) = 0.72, $p < 0.01$). A linear model of biological age was fit to chronological age and association with MDD was computed by comparing the number of MDD and HC subjects above and below the regression line.

In the current study, we create a biological age model from RNA-Seq gene expression using a multivariate LASSO penalized regression rather than an iterative univariate test, and we use age as a quantitative variable during the feature selection in linear regression, as opposed to using age quartiles as in Ref.[8], which allows our model to include more variation when estimating the age model. When later using chronological age as covariate for MDD association, we dichotomize chronological age. LASSO allows automatic feature selection of a multivariate linear regression model based on the cross-validated penalty hyperparameter optimization. We train the LASSO biological age model using an existing RNAseq dataset consisting of 157 individuals (78 with MDD and 79 healthy controls) [8, 9], and we use the residual of the LASSO model as an estimate of the gap between an individual's chronological age and their biological gene age. A positive gap indicates higher than average biological age or elevated aging compared to chronological age. This LASSO Gene Age Gap Estimate (L-GAGE) shows elevated biological aging in MDD subjects compared to HC, but the elevation is not statistically significant. However, when we dichotomize chronological age into older and younger, the interaction between MDD status and age is significant in L-GAGE model. Finally, we use machine learning feature selection to explore biological pathways that are significantly enriched for the gene sets identified as being associated with aging.

2. Materials and Methods

2.1. RNA-Seq Data

To test our biological age models, we use an extant RNA-Seq dataset [10]. The study was approved by the Western Institutional Review Board and conducted according to the principles expressed in the Declaration of Helsinki. The data consists of 78 MDD and 79 HC subjects (91 females and 66 males). Individuals with current symptoms of depression met DSM-IV-TR criteria for MDD based on the Structural Clinical Interview for DSM-IV-TR Axis I Disorders and an unstructured psychiatric interview. HC individuals had no personal or immediate family history of major psychiatric disorders. MDD participants were unmedicated for at least 3 weeks prior to study entry. Exclusion criteria included major medical or neurological illness, psychosis, traumatic brain injury, and a history of drug/alcohol abuse within 1 year. There is a higher female/male ratio for MDD (51/27) than HC (40/39), compatible with trends in the general population. The age distribution is slightly skewed towards younger individuals with age range from 18 to 55 (Fig. 1). The 8,923 genes in the RNA-Seq gene expression data are normalized by counts per million reads, which we then quantile normalize and log2 transform to stabilize variance. We removed genes with a low coefficient of variation (standard deviation divided by absolute mean). We chose a threshold of 0.045 to obtain 5,587 genes.

2.2. Gene Age Gap Estimate (GAGE) using RNA-Seq

We use LASSO for gene selection and modeling biological age, and then we use the residual of this model, which we call LASSO Gene-Age Gap Estimate (L-GAGE), for association testing with MDD. For the LASSO biological aging model, we build a full penalized regression model with all gene expression variables and with chronological age as the outcome variable. We include both MDD and HC samples in the age model, which was also the approach in Ref. [8]. Our biological age model is based on the non-zero coefficient genes from the lambda-1se

LASSO penalty (the largest λ for which the average cross-validation (CV) error is within one standard error of the minimum CV error). We compute the gap/residuals of the LASSO model between predicted biological age and chronological age (i.e., the L-GAGE score). Our goal is to use L-GAGE to test for increased biological age in MDD subjects (Fig. 2).

2.3. Relationship between gene age gap, chronological age, MDD and sex

It is important to consider adjustments for chronological age in biological age models because of regression to the mean as discussed for brain age models [11], but sex is also an important covariate for MDD. To further explore covariate effects, we add MDD x Age and MDD x Sex interactions for L-GAGE associations with MDD. We use the OLS model

$$LGAGE = \beta_0 + \beta_1 MDD + \beta_2 Z + \beta_3 (MDD * Z) + \varepsilon, \qquad \text{(Eq. 1)}$$

where Z represents the adjustment or interaction variable (Age or Sex). We focus on the effect of β_3, which represents how much the average L-GAGE of the MDD group changes for the Z=1 condition.

We consider two cases when age is used as a covariate with interactions (Z in Eq. 1): as continuous and as dichotomous with a threshold. To verify our choice of age threshold, we use a threshold regression model in the “chngpt” package in R [12]. We use this approach to check for possible nonlinear relationship between MDD and age and whether the effect of chronological age on MDD increases at some threshold point. The mean function of the threshold model is:

$$\eta = \alpha_1 + \alpha_2 z + \beta_1 I(x > e), \qquad \text{(Eq. 2)}$$

where x stands for chronological age, e is the age threshold and z are additional predictors. “I” is a step indicator function. The threshold is optimized using the exact criterion function with a logistic-based smooth function.

2.4. Feature selection, Gene-Age pathway Enrichment, and interpretable classifier

We use LASSO to create the gene-based residual age model, L-GAGE, but LASSO feature selection also results in a set of age-related genes. As a secondary analysis, we use LASSO and other feature selection methods to identify important age-related genes for pathway enrichment to understand the biological mechanisms of the age models. We use univariate linear regression, random forest (RF) regression, and nearest-neighbor projected distance regression (NPDR) [10] as feature selection methods. RF has the ability to find more complex models than LASSO and linear regression, but RF has limited ability to detect interactions [13], whereas NPDR has the ability to detect interaction effects [10]. For univariate feature selection, we use a linear model of individual genes with age, and we use a P-value threshold of 0.05 (uncorrected for improved pathway overlap). We use the standard NPDR with an adjusted P-value threshold of 0.05 FDR, and we use the LASSO penalized NPDR. For NPDR, we use the imbalanced k-nearest-neighbor value (k=47) that approximates the 0.5 standard deviation of the hyper-radius [10]. We use permutation variable importance with RF. We use the Reactome Pathway database in MSigDB [14, 15] for biological pathway enrichment of age related genes. For additional interpretation of the gene-age prediction of MDD along with consideration for other covariates, we train a decision tree to predict MDD based on L-GAGE, chronological age, and sex. Decision trees have high variance, but they are useful for interpreting the relationships between covariates.

3. Results and Discussion

3.1. Testing Association of Gene Age L-GAGE with MDD.

We test for association of the LASSO Gene Age Gap Estimate (L-GAGE) score with MDD status. L-GAGE is the residual from a LASSO gene expression model of chronological age. The LASSO model uses the cross-validation tuned lambda-1se value (*$\lambda = 1.636048$*), which is the largest λ at which the mean-squared error (MSE) is within one standard error of the minimum MSE. The residuals are constant, and heteroscedasticity is not present based on the Non-constant Variance Score Test. The penalty results in a multivariate linear model of age with 22 genes and a Spearman Correlation Coefficient (SCC) with chronological age of 0.77 (Fig. 2). Counting the number of HC or MDD above or below the regression line (Fig. 2), we find that the biological age is greater in MDD subjects than HC (HC – 45 (56.96 %) below, 34 (43.037%) above, MDD 35 (44.87%) below, 43 (55.128%) above). The P-value of the Chi-squared test of GAGE sign (above or below the line) for MDD is not significant (0.1753). The greater L-GAGE in MDD versus HC can be seen in L-GAGE density (Fig. 3A). The L-GAGE distribution for males and females is very similar (Fig. 3B). While L-GAGE is greater in MDD than HC subjects, we do not find a statistically significant replication of the effect found in Ref. [8]. However, we do see a suggestive difference with an effect size similar to what they found. Using the same genes as their model also does not replicate.

3.2 Testing MDD-Age interaction for L-GAGE association model.

We test for the effect of L-GAGE on MDD by introducing an MDD-Age interaction term (Eq. 1). Dichotomizing age at threshold 40, MDD alone is not significant, but we find a statistically significant effect of the interaction between MDD and Age 40 on L-GAGE (Table 1, Fig. 4). For individuals younger than 40, L-GAGE shows very little difference between MDD and HC, but for older individuals, there is greater biological aging (L-GAGE) for the MDD versus HC group (Fig. 4 and Table 1). Age alone is also statistically significant (Table 1). These age effects remain significant when we add sex as a covariate (Table 1B), but sex is not significant (Table 1B and Table 2).

The MDD-Age interaction and the MDD term (Eq. 1) do not have a significant effect on L-GAGE when age is treated as a continuous variable (MDD P-value = 0.364, Age P-value = 0.316, MDD*Age P-value = 0.197). Also, there is no direct statistical association between MDD and age and between MDD and sex (Two Sample T-test of MDD and Chronological age: P-value = 0.167; Chi-squared-test of MDD and sex: P-value = 0.08716). To further support our choice of age threshold, we use a threshold regression (Eq. 2). The change point for age in relation to MDD is estimated to be 39 years (Fig. 5). Combined with the third quartile being age 41, the threshold regression suggests that age 40 is a suitable cutoff point for dividing the subjects into two age groups.

Additional support for the age-40 threshold can be seen in the decision tree for predicting MDD (Fig. 6), where age with threshold 39.5 is the second important split variable, following L-GAGE. The decision tree also suggests interaction effects, where the effect of L-GAGE on MDD is conditioned on chronological age. If L-GAGE (top node) is below a threshold, subjects tend to be HC. If the L-GAGE is below the threshold and chronological age is above 39.5 (i.e., an interaction), subjects tend to be MDD. However, for chronological age less than 39.5, the prediction of MDD is more complex (Fig. 6). We note that this decision tree was trained on the full dataset to maximize power, but it is instructional for interpretation.

A subset of our subjects (136 out of 157) have anti-CMV (human cytomegalovirus) IgG antibody data. Of the 136 samples, 70 are CMV seropositive and 66 CMV seronegative CMV. Although the P-value is not significant (0.097), we find that the mean biological age gap (L-GAGE) is higher in CMV positive subjects compared to CMV negative (Fig. 7A). For the subset of subjects with both CMV data and MDD status data, there are 75 HC and 61 MDD and 83 female and 54 male. While CMV positive subjects tend to have an elevated biological age, the effect is not MDD or sex specific (Fig. 7B and 7C). In other words, being CMV positive elevates gene age regardless of MDD/HC status or sex.

3.3 Characterizing Age-Associated Genes

The LASSO regression used in L-GAGE selected 22 age genes with non-zero coefficients (Table 3). We broaden the characterization of age related genes in our MDD data through pathway enrichment from statistical and machine learning feature selection methods linear regression, RF, and nearest-neighbor projected distance regression (NDPR) [10]. Across all feature selection methods, the four common age genes are NAA20 (N-alpha-acetyltransferase 20), CCNE1 (Cyclin E1), and SESTD1 (SET domain containing protein 1A), and TAF9 (TATA-box-binding protein associated factor 9). Using the feature selection gene sets and the Reactome database, we find enrichment for Infectious Disease, Adaptive Immune System, and SARS-CoV-2 Infection pathways (Tables 5 and 6). SARS-CoV-2 can cause neurological complications, and a recent study showed that differentially expressed genes for COVID infection overlap with many gene associations for neuropsychiatric disorders including depression [16].

Conclusion

We presented a procedure for creating an expression-based biological age model using LASSO penalized regression, and we explored the association of the residual, or the LASSO-based Gene Age Gap Estimate (L-GAGE) on MDD while adjusting for chronological age and sex. We found increased biological aging based on L-GAGE in MDD versus HC subjects with an effect size similar to a previous study [8], but the difference was not statistically significant. Larger sample sizes are needed to further test this effect. We found a statistically significant MDD-Age interaction for L-GAGE when age is dichotomized with threshold 40 years. We used multiple statistical criteria for choosing this threshold. This finding could indicate an effect of lifetime number of MDD episodes on biological aging that is not detectible until middle-age. The interaction effect remained significant when adjusting for chronological age and sex, and we reiterate the importance of including age in L-GAGE association tests to avoid confounding due to regression to the mean [11].

We explored the top age-associated genes with different feature selection methods, and we identified a consensus set of genes, CCNE1, NAA20, SESTD1, and TAF9 that have been associated with aging, senescence, and infectious disease. In a study of Lung Adenocarcinoma, CCNE1 gene expression was found to be correlated with patients' age [17], and NAA20 and SETD1A are involved in senescence, which is related to aging and age-related diseases. It was shown that depletion of NAA20 in non-transformed mammal cells led to senescence [18], and in another study knockdown of SETD1A triggered cellular senescence. [19]. TAF9 cross-reactivity was shown to be associated with immunity to CMV in the context of autoimmune disease [20]. Recall, we found that CMV positive status is associated with elevated biological age based on L-GAGE. Pathway enrichment of the broader set of age genes selected by linear regression, random forest, and NPDR resulted in the detection of Infectious Disease, Adaptive Immunity, and SARS-CoV Infection pathways. As noted in Ref. [8], evaluating PBMC transcription can increase the risk for false positive immune pathways.

This study contributes a new approach to estimating biological aging and contributes to the evidence for the role of aging and inflammation in depression. Future studies are needed with broader age ranges, more uniform age distributions, large sample sizes, and utilization of MDD age-of-onset and number of depressive episodes. Future gene age models may help identify individuals that need different treatment or management for depression due to an increase in their relative biological age.

Research data for this article

Data and code for this research will be available at https://github.com/insilico/GeneAgeMDD.

Funding

BAM and JS received support from the National Institute of Mental Health (R01MH098099).

Reference

1. Ford BN, Savitz J: **Depression, aging, and immunity: implications for COVID-19 vaccine immunogenicity**. *Immunity & Ageing* 2022, **19**(1):32.
2. Darrow SM, Verhoeven JE, Révész D, Lindqvist D, Penninx BW, Delucchi KL, Wolkowitz OM, Mathews CA: **The Association Between Psychiatric Disorders and Telomere Length: A Meta-Analysis Involving 14,827 Persons**. *Psychosom Med* 2016, **78**(7):776-787.
3. Ridout KK, Ridout SJ, Price LH, Sen S, Tyrka AR: **Depression and telomere length: A meta-analysis**. *J Affect Disord* 2016, **191**:237-247.
4. Ait Tayeb AEK, Poinsignon V, Chappell K, Bouligand J, Becquemont L, Verstuyft C: **Major Depressive Disorder and Oxidative Stress: A Review of Peripheral and Genetic Biomarkers According to Clinical Characteristics and Disease Stages**. *Antioxidants* 2023, **12**(4):942.
5. Raison CL, Capuron L, Miller AH: **Cytokines sing the blues: inflammation and the pathogenesis of depression**. *Trends Immunol* 2006, **27**(1):24-31.
6. Protsenko E, Yang R, Nier B, Reus V, Hammamieh R, Rampersaud R, Wu GWY, Hough CM, Epel E, Prather AA *et al*: **"GrimAge," an epigenetic predictor of mortality, is accelerated in major depressive disorder**. *Translational Psychiatry* 2021, **11**(1):193.
7. Han LKM, Dinga R, Hahn T, Ching CRK, Eyler LT, Aftanas L, Aghajani M, Aleman A, Baune BT, Berger K *et al*: **Brain aging in major depressive disorder: results from the ENIGMA major depressive disorder working group**. *Mol Psychiatry* 2021, **26**(9):5124-5139.
8. Cole JJ, McColl A, Shaw R, Lynall ME, Cowen PJ, de Boer P, Drevets WC, Harrison N, Pariante C, Pointon L *et al*: **No evidence for differential gene expression in major depressive disorder PBMCs, but robust evidence of elevated biological ageing**. *Transl Psychiatry* 2021, **11**(1):404.
9. Li YJ, Kresock E, Kuplicki R, Savitz J, McKinney BA: **Differential expression of MDGA1 in major depressive disorder**. *Brain Behav Immun Health* 2022, **26**:100534.
10. Le TT, Dawkins BA, McKinney BA: **Nearest-neighbor Projected-Distance Regression (NPDR) for detecting network interactions with adjustments for multiple tests and confounding**. *Bioinformatics* 2020, **36**(9):2770-2777.
11. Le TT, Kuplicki RT, McKinney BA, Yeh HW, Thompson WK, Paulus MP, Tulsa I: **A Nonlinear Simulation Framework Supports Adjusting for Age When Analyzing BrainAGE**. *Front Aging Neurosci* 2018, **10**:317.
12. Fong Y, Huang Y, Gilbert PB, Permar SR: **chngpt: threshold regression model estimation and inference**. *BMC Bioinformatics* 2017, **18**(1):454.
13. McKinney BA, Crowe JE, Guo J, Tian D: **Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis**. *PLoS Genet* 2009, **5**(3):e1000432.
14. [https://www.gsea-msigdb.org/gsea/msigdb/index.jsp]
15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545-15550.
16. Quincozes-Santos A, Rosa RL, Tureta EF, Bobermin LD, Berger M, Guimaraes JA, Santi L, Beys-da-Silva WO: **COVID-19 impacts the expression of molecular markers**

associated with neuropsychiatric disorders. *Brain Behav Immun Health* 2021, **11**:100196.

17. Ullah MA, Farzana M, Islam MS, Moni R, Zohora US, Rahman MS: **Identification of the prognostic and therapeutic values of cyclin E1 (CCNE1) gene expression in Lung Adenocarcinoma and Lung Squamous Cell Carcinoma: A database mining approach**. *Heliyon* 2022, **8**(9):e10367.
18. Elurbide J, Carte B, Guedes J, Aldabe R: **NatB Catalytic Subunit Depletion Disrupts DNA Replication Initiation Leading to Senescence in MEFs**. *Int J Mol Sci* 2023, **24**(10).
19. Tajima K, Matsuda S, Yae T, Drapkin BJ, Morris R, Boukhali M, Niederhoffer K, Comaills V, Dubash T, Nieman L *et al*: **SETD1A protects from senescence through regulation of the mitotic gene expression program**. *Nature Communications* 2019, **10**(1):2854.
20. Chen YF, Hsieh AH, Wang LC, Yu KH, Kuo CF: **Cytomegalovirus-Associated Autoantibody against TAF9 Protein in Patients with Systemic Lupus Erythematosus**. *J Clin Med* 2021, **10**(16).

Figures

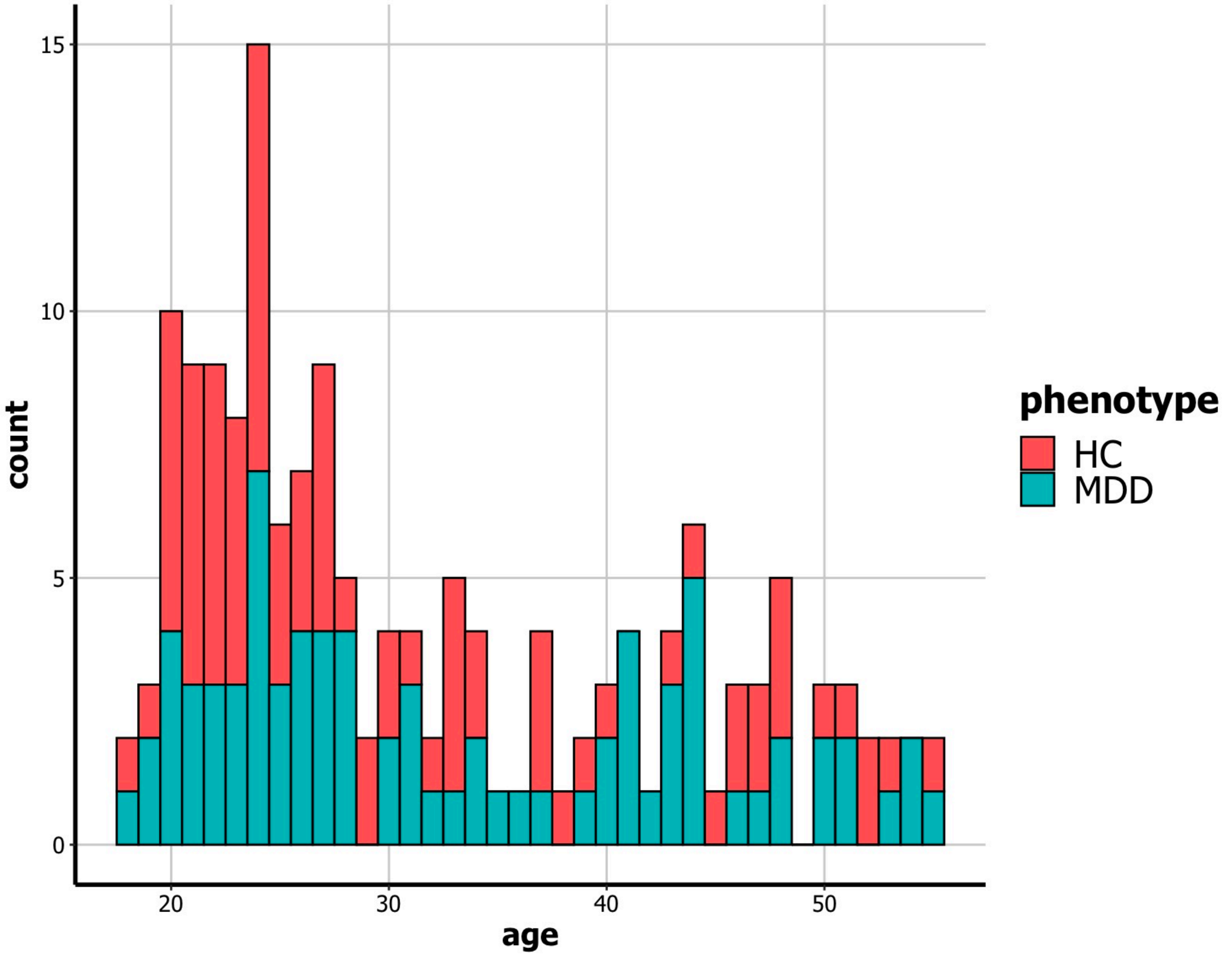

Figure 1. Histogram of chronological ages with a bin size of 1: Bars are separated by Healthy Control (HC, red) and major depressive disorder (MDD, blue). There are more younger subjects in the dataset with the same age, especially from age 20~28. For example, there are 15 subjects that are 24 years old. Chronological age is not associated with MDD versus HC (T-test P-value 0.167).

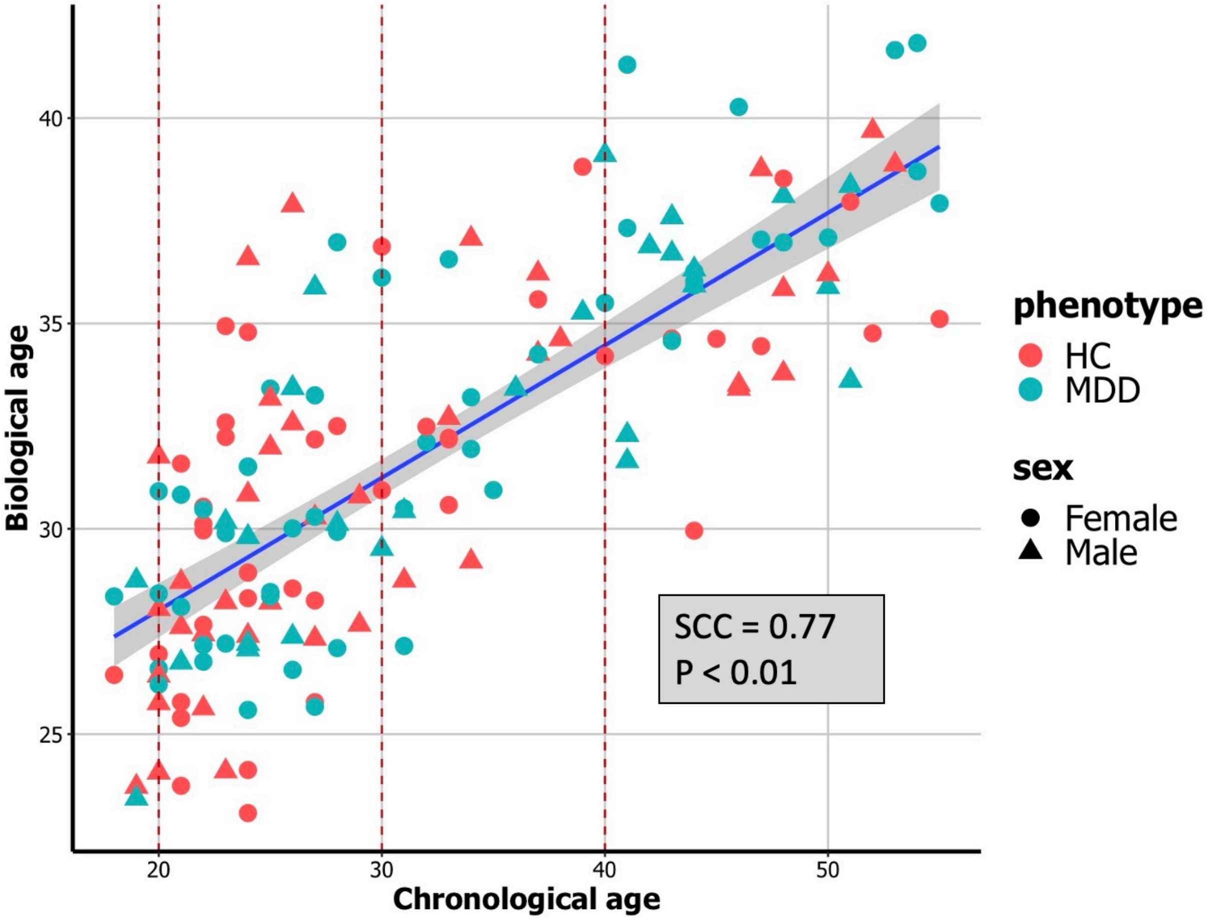

Figure 2. Scatter plot with regression line of biological age and chronological age: Biological age model is based on LASSO regression and the residual is later used for LASSO Gene Age Gap Estimate (L-GAGE). The points are colored by MDD (blue) and HC (red)**.** The points are shaped by Female (circle) and Male (triangle)**.** Spearman Correlation Coefficient (SCC = 0.77, slope P-value < 0.01)**.**

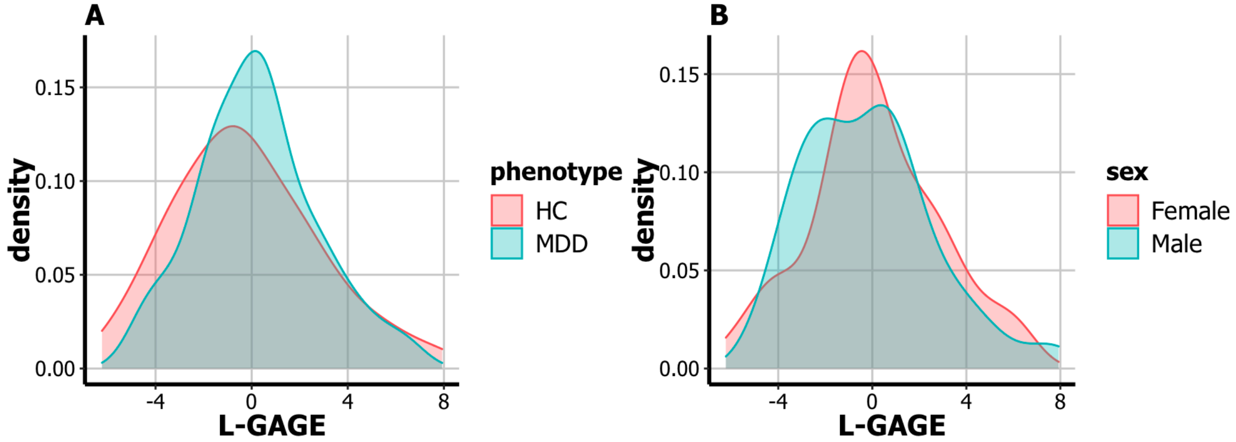

Figure 3. Density plots of the LASSO based Gene Age Gap Estimate (L-GAGE) separated by MDD (A) and sex (B). A positive gene-age residual (x-axis) indicates a sample above the gene age regression line and negative below. **A**. Biological age relative to chronological age (L-GAGE) is greater in MDD patients than in HC. **B**. The L-GAGE difference between males and females is less pronounced.

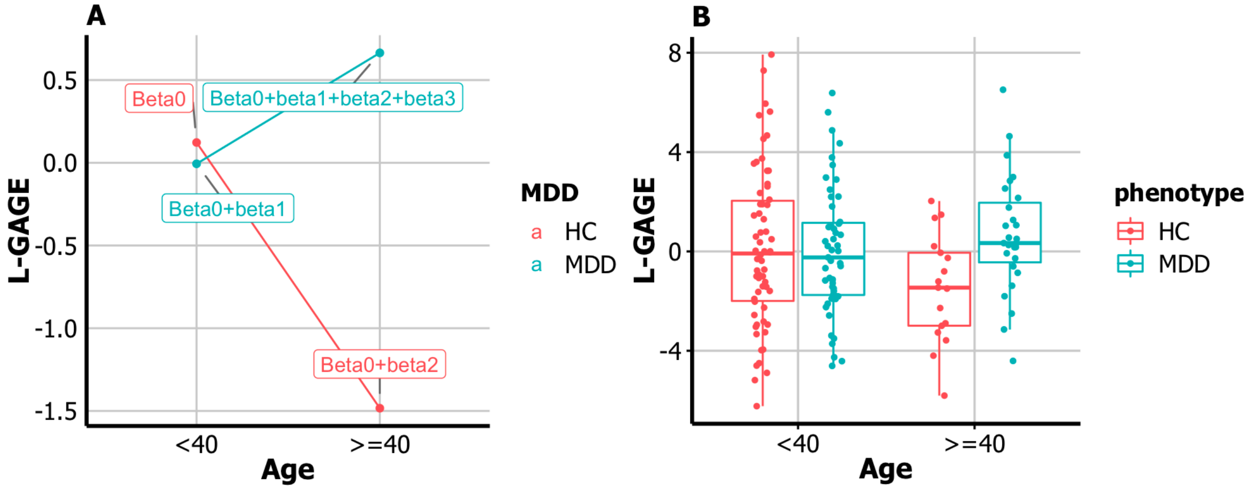

Figure 4. MDD x Age interaction for L-GAGE with age 40 threshold. A. The average L-GAGE for people older than 40 with MDD is higher than the L-GAGE value for people younger than 40 with MDD (blue line), whereas in the HC group the average L-GAGE is lower for people older than 40 than for people younger than 40 (red line). **B.** For individuals younger than 40, L-GAGE shows very little difference between MDD and HC. For older individuals, there is greater biological aging (L-GAGE) for the MDD versus HC group. The L-GAGE association with MDD is still significant when adjusted by age and sex.

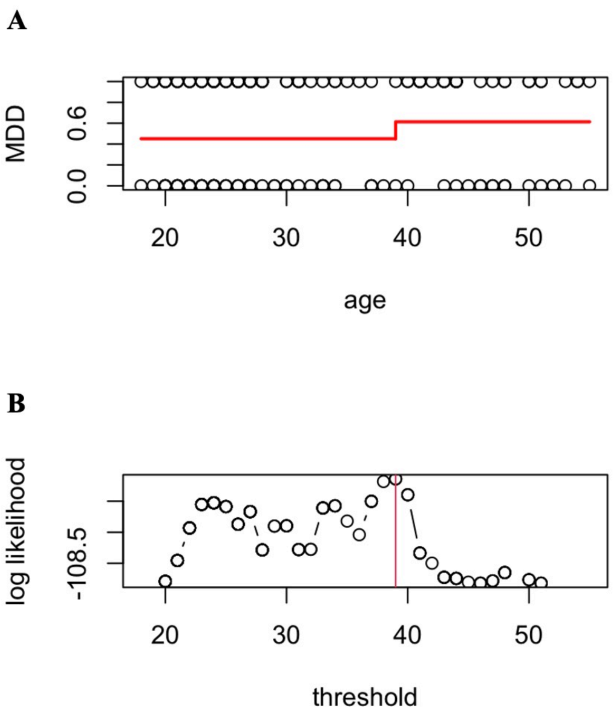

Figure 5. Effect of Chronological Age on MDD Determined by Threshold Regression Model. **A.** Threshold regression (Eq. 2) shows the nonlinear relationship between MDD and chronological age. The prediction indicates an increase in MDD up to the age of 39, which is identified as the change point by the model. **B.** The likelihood analysis of the threshold regression model also indicates that age 39 is the optimal threshold, having the highest model likelihood.

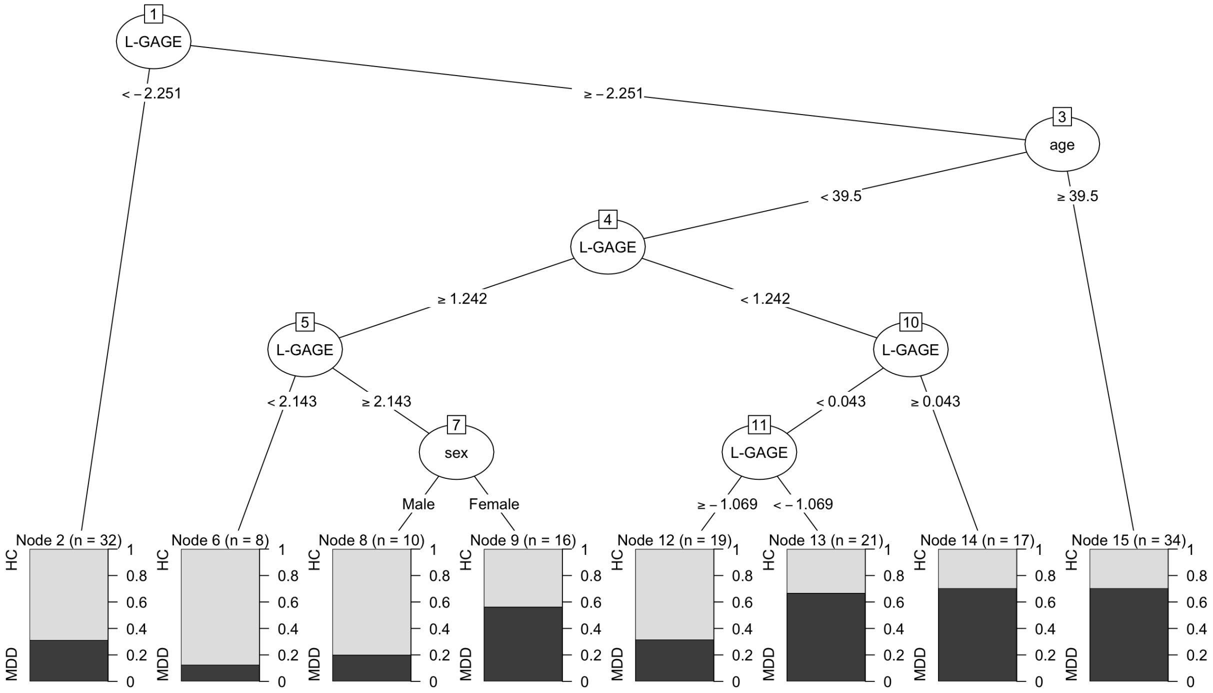

Figure 6. Gene age decision tree for MDD with covariates. For added interpretation, we train a decision tree on all samples to predict MDD. The model identifies the gene age residual L-GAGE as the most important predictor, with chronological age being the second most significant factor. In the first split, if the gene age gap is low, L-GAGE < -2.251 (Node 1), there is high probability for a subject to be HC (Node 2). If the gene age gap is higher, L-GAGE ≥ -2.251, the model becomes more complex and initially depends on chronological age with split 39.5 years (Node 3). If L-GAGE is high and Age ≥ 39.5, then there is a high probability a subject is MDD (Node 15). When Age < 39.5, the model again becomes dependent on L-GAGE, and at a certain split, females exhibit a higher probability of MDD compared to males (Nodes 8 and 9).

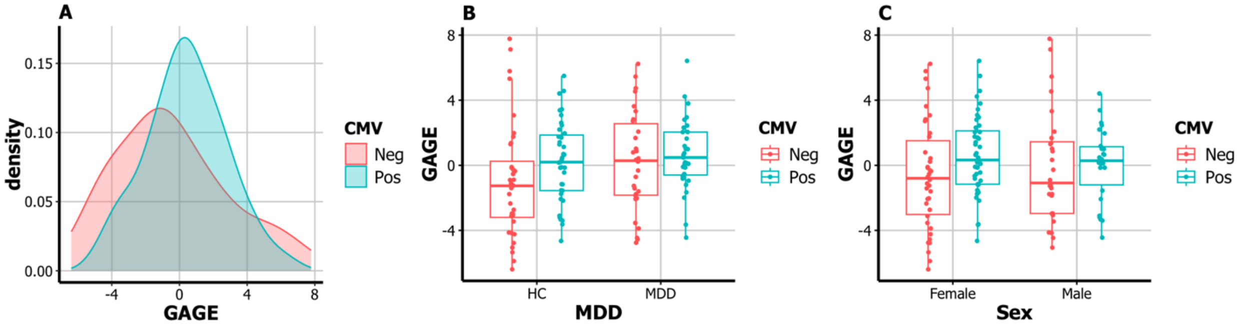

Figure 7. Distribution of Gene Age Gap Estimate (GAGE) conditioned on positive/negative cytomegalovirus (CMV) status. A. Mean biological age (GAGE) relative to chronological age is greater in CMV positive subjects (blue) than in CMV negative (red). B. Healthy controls (HC) that are CMV positive (blue) have a higher GAGE than CMV negative subjects. The MDD-CMV+ subjects also have a slightly higher GAGE than MDD-CMV- subjects, but the difference in GAGE for MDD subjects based on CMV status is very small. C. Similarly, mean biological age relative to chronological age based on GAGE increases with positive CMV status for both females and males.

Tables

Table 1. LASSO Gene-age-gap estimate (L-GAGE) association with MDD and dichotomized age interaction. A. Based on the ordinary least squares model (Eq. 1 with Z=Age), where chronological age is dichotomized with threshold is Age>=40 and Age<40, the MDD x Age interaction is significant. Biological age (L-GAGE)) is similar for MDD and HC when Age<40, but when the chronological age is higher than 40, biological age is significantly greater in MDD individuals than HC. **B**. The MDD x Age interaction remains significant when Sex is added as a covariate.

A

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1225	0.3495	0.35	0.7265
MDD	-0.1286	0.5203	-0.247	0.8052
Age40	-1.606	0.7535	-2.131	0.0347*
MDD*Age40	2.2764	0.9984	2.28	0.024*
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

B

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2644	0.4172	0.634	0.5273
MDD	-0.1870	0.5296	-0.353	0.7246
SexMale	-0.2838	0.4534	-0.626	0.5324
Age40	-1.6144	0.7551	-2.138	0.0341*
MDD*Age40	2.3274	1.0038	2.319	0.0217*
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 2. Gene-age-gap regression with MDD-sex interaction with Female and Male. Based on the ordinary least squares model (Eq. 1 with Z=Male/Female instead of age), L-GAGE score of MDD in males is slightly lower than the L-GAGE score of MDD in females, but the interaction term MDD*Male is not statistically significant.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.24481	0.44221	-0.554	0.581
MDD	0.64637	0.5907	1.094	0.276
Male	0.04395	0.62938	0.07	0.944
MDD* Male	-0.55121	0.91608	-0.602	0.548
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Table 3. Age associated genes selected by LASSO. Multivariate coefficients are shown that survived LASSO penalty. Negative coefficients (left columns) indicate higher expression of the gene tends to occur with younger age. Positive coefficients (right columns) indicate higher expression of the gene tends to occur in older individuals. These genes are used in the gene age model and the L-GAGE residual.

Down Regulated with Increasing Age		**Up Regulated with Increasing Age**	
Gene	**Coefficient**	**Gene**	**Coefficient**
NAA20	-6.7070152	CCNE1	14.2689027
ZNF347	-2.9514771	SESTD1	8.8624231
PRMT6	-2.4559818	ZNF334	2.4209761
WDR13	-1.7979357	ANTXRL	2.0255277
DDX19B	-1.3737037	DTD2	1.8502139
TAF9	-1.2672137	CYTH3	1.5349361
ADSS	-1.1724134	DYRK1A	1.2905045
TGFBR3	-1.0316785	HTATSF1	1.078388
SMYD5	-0.8454683	SFXN4	0.7870119
CISD1	-0.6212633	UBE2F-SCLY	0.2252943
TGIF2-C20orf24	-0.5057642		

Table 4. Age associated genes selected by linear regression with adjusted P-value 0.05 FDR. Negative coefficients (left columns) indicate higher expression of the gene tends to occur with younger age. Positive coefficients (right columns) indicate that higher expression of the gene tends to occur in older individuals. These genes are shown for comparison but not used in the gene age model.

Down Regulated with Increasing Age				**Up Regulated with Increasing Age**			
Gene	**Coefficient**	**P-value**	**Adjusted P-value**	**Gene**	**Coefficient**	**P-value**	**Adjusted P-value**
NAA20	-16.1918	8.86E-08	0.0005	CCNE1	42.8022	5.59E-07	0.0013
CIART	-22.7969	6.87E-07	0.0013	SESTD1	12.2045	1.19E-05	0.0111
TAF9	-21.2804	2.85E-06	0.0040	ITGB1BP1	10.7847	2.46E-05	0.0197
MLXIPL	-20.0949	4.55E-06	0.0051	ANTXRL	13.8739	4.23E-05	0.0295
TGFBR3	-17.7019	7.91E-05	0.0491				

Table 5. MSigDB Reactome results of the age genes selected by linear regression. We collect the 464 age associated genes with P-value lower than 0.05 (not adjusted for better pathway

detection) and query MSigDB Reactome database for pathway enrichment. Notably, these age associated genes are enriched for infectious disease and SARS-CoV Infections pathways.

Gene Set Name	Genes in Gene Set (K)	Description	Genes in Overlap (k)	k/K	p-value
REACTOME_RNA_POLYMERASE_II_TRANSCRIPTION	1393	RNA Polymerase II Transcription	46	0.0330	3.84E-11
REACTOME_POST_TRANSLATIONAL_PROTEIN_MODIFICATION	1442	Post-translational protein modification	44	0.0305	1.21E-09
REACTOME_METABOLISM_OF_RNA	714	Metabolism of RNA	29	0.0406	2.05E-09
REACTOME_TRANSCRIPTIONAL_REGULATION_BY_TP53	363	Transcriptional Regulation by TP53	20	0.0551	4.74E-09
REACTOME_INFECTIOUS_DISEASE	1019	Infectious disease	33	0.0324	3.95E-08
REACTOME_MEMBRANE_TRAFFICKING	629	Membrane Trafficking	23	0.0366	6.11E-07
REACTOME_METABOLISM_OF_LIPIDS	742	Metabolism of lipids	25	0.0337	8.86E-07
REACTOME_SUMOYLATION	187	SUMOylation	12	0.0642	1.19E-06
REACTOME_SARS_COV_INFECTIONS	471	SARS-CoV Infections	19	0.0403	1.4E-06
REACTOME_VESICLE_MEDIATED_TRANSPORT	724	Vesicle-mediated transport	23	0.0318	6.34E-06

Table 6. MSigDB Reactome results of the 145 age genes selected by nearest-neighbor projected distance regression (NPDR) with LASSO penalty.

Gene Set Name	Genes in Gene Set (K)	Description	Genes in Overlap (k)	k/K	p-value
REACTOME_NEF_MEDIATES_DOWN_MODULATION_OF_CELL_SURFACE_RECEPTORS_BY_RECRUITING_THEM_TO_CLATHRIN_ADAPTERS	21	Nef-mediates down modulation of cell surface receptors by recruiting them to clathrin adapters	4	0.1905	7.44E-07
REACTOME_NEF_MEDIATED_CD4_DOWN_REGULATION	9	Nef Mediated CD4 Down-regulation	3	0.3333	3.22E-06
REACTOME_RNA_POLYMERASE_II_TRANSCRIPTION	1393	RNA Polymerase II Transcription	16	0.0115	2.38E-05
REACTOME_LDL_CLEARANCE	19	LDL clearance	3	0.1579	3.62E-05
REACTOME_TRANSCRIPTIONAL_REGULATION_BY_TP53	363	Transcriptional Regulation by TP53	8	0.022	3.78E-05
REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION	126	MHC class II antigen presentation	5	0.0397	7.56E-05
REACTOME_ADAPTIVE_IMMUNE_SYSTEM	829	Adaptive Immune System	11	0.0133	1.38E-04
REACTOME_TRAFFICKING_OF_AMPA_RECEPTORS	31	Trafficking of AMPA receptors	3	0.0968	1.63E-04
REACTOME_TP53_REGULATES_METABOLIC_GENES	87	TP53 Regulates Metabolic Genes	4	0.046	2.32E-04