Age-Related Hippocampal and Striatal Gene Expression

Pam Rivière
University of California, San Diego

Dillan Cellier
University of California, San Diego

Eena Kosik
University of California, San Diego

Mia Borzello
University of California, San Diego

1 INTRODUCTION

Age-related cognitive decline has been shown to affect 32% of adults aged 71 to 89 years (Machulda et al., 2013). As living conditions improve and life expectancy increases across the globe, it will become necessary to address age-related cognitive decline with a higher degree of sophistication than is currently available. Specifically, mitigating the adverse effects of aging will require a mechanistic understanding of the cellular and molecular machinery that underlies the brain's functional decline.

Genes encode proteins that are essential to (1) meet the common metabolic needs across all neurons, as well as (2) equip neurons with unique intrinsic properties (e.g. idiosyncratic distributions of ion channels, etc). This latter category of genes are differentially expressed across distinct neuron types, and more broadly, across distinct brain regions. Region-specific genetic programs may be affected differently over the course of an organism's lifetime.

Two regions of the brain, the hippocampus and striatum, are positioned to play a critical role in age-related cognitive decline. Cell death, neuroinflammation, a reduction in neurogenesis, and changing epigenetics in the hippocampus have all been implicated in age-associated cognitive decline (Bettio et al., 2017). Similarly, volume reduction in the striatum, and specifically in the putamen and nucleus accumbens, has been found to correspond to more severe cognitive decline (de Jong et al, 2012). It is therefore important to more closely examine the role of individual genes and gene expression in these brain regions as they relate to cognitive processes known to decline with age.

Technical advances in the field of genetics now make it possible to investigate age-related changes in gene expression across neural regions. In the work that follows, we leveraged the freely available AGEMAP dataset (Zahn et al., 2007), which documents gene expression levels across multiple brain regions from differently-aged cohorts of mice. To assess which genes' expression levels systematically varied with age, we performed linear regression with coefficient fits subject to Lasso regularization. In these models, the age of mice subjects was regressed against a linear combination of gene expression levels in either hippocampus or striatum. This procedure resulted in a sparse subset of genes whose expression levels covaried with age in each region.

Prior to fitting the models, we expected the selected hippocampal subset of genes to predominantly include those involved in the production of proteins integral to synaptic plasticity, given that compromised synaptic potentiation is a strong predictor of memory deficits in hippocampus (Foster, 1999). In the striatum, we anticipated that selected genes would be related to the expression of a class of potassium ion channels (KCQN) whose currents crucially modulate medium spiny neuron activity, the predominant principal neuron type in this region (Shen et al., 2005; McCarthy et al., 2011). Instead, we found that the selected subset of genes overwhelmingly related to basic metabolic functions and structural properties that might, in principle, be relatively common across neurons in distinct regions. This pattern of results suggests that the basic fabric of neuronal morphology and homeostasis should not be overlooked when considering clinical interventions against age-related cognitive decline.

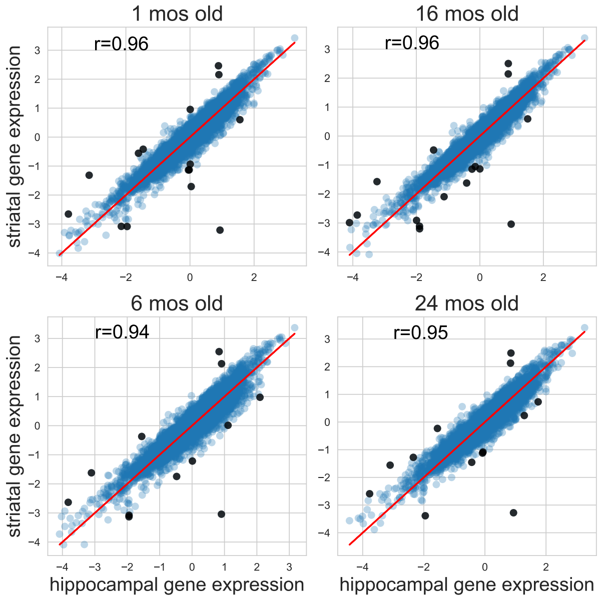

Figure 1: ***Differential gene expression across brain regions, by age group.*** *Each panel corresponds to a mouse age group (mos = months). Blue dots correspond to the mean expression of a given gene within striatum against the same gene's mean expression in hippocampus, for a given age group. Black dots mark genes whose absolute difference in hippocampal versus striatal expression was 5 standard deviations above the average of the absolute differences across genes, for a given age group. Pearson rs correspond to the correlation in gene expression between regions.*

2 METHODS

2.1 Dataset

Data are drawn from the AGEMAP (Atlas of Gene Expression in Mouse Aging Project) database (Zahn et al., 2007), which generates mRNA transcript profiles across 8,932 genes and 16 mouse tissues at four times during aging (1, 6, 16, and 24 months). Unlike other

Table 1: *Genes that exhibit differential expression across hippocampus and striatum, computed within age group. Some genes were differentially expressed only in some age groups and not others, but they were still included as predictors in Lasso regression. Dashes indicate that a gene was not differentially expressed in that age group.*

1 mos. old	6 mos. old	16 mos. old	24 mos. old
–	–	–	Wdr22
Tsg101	Tsg101	Tsg101	Tsg101
Suv39h2	Suv39h2	Suv39h2	Suv39h2
Snrpb2	Snrpb2	Snrpb2	Snrpb2
–	–	–	Scn3a
Skil	Skil	Skil	–
–	–	Sdbcag84	–
Plekhh1	Plekhh1	Plekhh1	Plekhh1
Npm3	Npm3	Npm3	–
–	–	Msh6	–
–	Mm.381201	–	–
–	Hyal1	–	–
Jph1	–	–	–
Hspd1	–	Hspd1	Hspd1
Gga2	–	–	–
–	–	Dok4	–
–	–	–	Ccm2
C1qr1	–	–	–
BC002230	–	BC002230	–
Anxa1	Anxa1	Anxa1	Anxa1
–	–	A930025J12Rik	–
–	9930021J03Rik	9930021J03Rik	–
–	–	–	8430437G11Rik
5330435L01Rik	5330435L01Rik	5330435L01Rik	5330435L01Rik
2700038I16Rik	–	–	–
–	–	–	2610509G12Rik
2410005K20Rik	2410005K20Rik	2410005K20Rik	2410005K20Rik

studies that compare genetic expression across age using different methodologies, AGEMAP is a highly standardized study using an identical gene array platform and experimental protocol for the same set of mice.

Specifically, we used genetic expression data from the hippocampus and striatum, which contains 10 samples for each gene, for each age group and is split evenly between genders (5 male and 5 female mice), totaling N=40 observations per gene.

2.2 Data Preprocessing

Genetic expression data was combined with the mouse meta data (mouse name, age, and sex) for each brain region (hippocampus and striatum). Any genes that were not available for both brain regions were dropped, resulting in a final subset of 6, 840 genes.

Of these 6, 840 genes, a large fraction exhibited similar expression levels between hippocampus and striatum within each age group (**Figure 1**). We inferred that overlapping genes were essential to the metabolic and structural features that should be commonly found across cells, irrespective of regional provenance in the brain. These overlapping genes might thus be less likely to subserve unique roles that could account for region-specific functional decline.

We consequently chose to analyze a smaller subset of genes that might be differentially expressed across regions. Here, a differentially expressed gene is one whose absolute difference in expression across hippocampus and striatum is five standard deviations larger than the mean absolute difference in expression values of all 6, 840 genes:

$$\Delta(x_{i,g}^{h-s}) = \mid (\mid x_{i,g}^{h} \mid - \mid x_{i,g}^{s} \mid) \mid \qquad (1)$$

where $\Delta(x_{i,g}^{h-s})$ represents the difference in hippocampal and striatal expression level x for the i^{th} gene in the g^{th} age group, $\mid x_{i,g}^{h} \mid$ represents the absolute $\log_2$ and z-scored hippocampal (h) expression level x for the i^{th} gene. The last term follows the same notation, but for the striatal (s) expression level of that gene. Expression levels are transformed such that,

$$x_{i,g} = \log_2 \chi_{i,g} \qquad (2)$$

where $\chi_{i,g}$ corresponds to the original, untransformed expression level for the i^{th} gene in the g^{th} age group prior to z-scoring (detailed z-scoring procedure below).

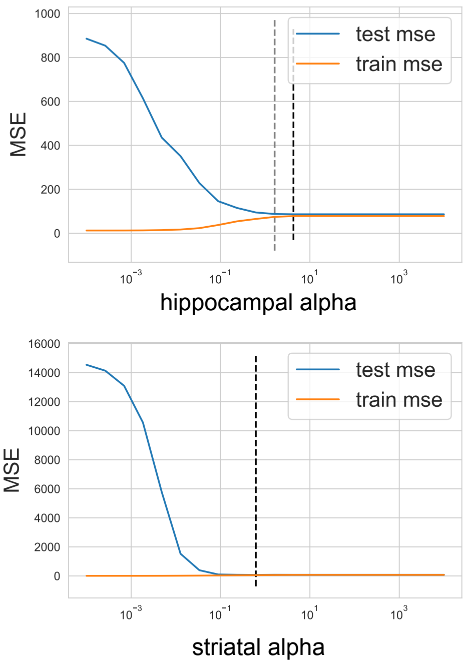

Figure 2: ***Optimization of regularization strength (alpha) for each region's model.*** *We separately fitted 20 different models for each region (hippocampus and striatum), where each model was subject to a different magnitude of regularization strength. For each model, we computed the mean squared error (MSE) on held-out test data (as well as training data, for comparison). We selected the alpha that minimized the average MSE across 5 held-out test sets. Black dotted line marks optimal alpha (hippocampal alpha = 4.28, striatal alpha = 0.62). Grey dotted line in upper panel marks the hippocampal alpha immediately preceding optimal alpha (= 1.62).*

Gene expression levels are heavily skewed on a linear scale because genes with lower expression tend to exhibit expression levels between 0 to 1, while higher-expressed genes exhibit values that range from 1 to ∞. Therefore, expression levels are *log*2-transformed to better represent this distribution, and subsequently z-scored to standardize the values across the dataset.

Z-scoring proceeds as follows: z-scores are calculated by subtracting the overall average gene intensity (within a single experiment) from the raw intensity data for each gene, and dividing that result by the standard deviation of all of the measured intensities. Each z-scored expression level is thus standardized relative to the distribution of expression values in the entire dataset drawn from a given experiment, allowing for expression level comparisons across animals, and across regions.

We computed expression differences in each age group separately, and we selected genes whose expression levels exceeded 5 standard deviations beyond the mean $\Delta\overline{x}_g^{h-s}$ across all genes in at least one age group. This pre-selection yielded 27 genes that served as model predictors (**Table 1**).

2.3 Lasso-Regularized Linear Regression

We performed two separate linear regressions, and applied the Lasso (Least Absolute Shrinkage and Selection Operator) penalty to each model's respective objective function. The first regression leverages hippocampal gene expression data to predict age, while the second leverages striatal gene expression data. The linear model is defined below:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pn} \tag{3}$$

where y_i is the age of the ith observation, n is the maximum number of observations, and p is the maximum number of predictors, which in this case, is genes. Then we apply the Lasso shrinkage to these models, which adds a regularization term that estimates a sparse subset of non-zero coefficients. We find the coefficient vector β that minimizes the following loss function:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad \text{s.t.} \quad \alpha\sum_{j=1}^{p}|\beta_j| \tag{4}$$

where the ordinary least-squares (OLS) term is subject to the L1 constraint $\sum_{j=1}^{p}|\beta_j|$, whose efficacy is parameterized by the regularization strength constant α.

To find the α that minimized the mean squared error of our predictions relative to held-out data (test MSE), we performed k-fold cross-validation, (using the "Kfold" function from the SciKit-Learn library). We used a k value of 5, to split the dataset into 5 consecutive train-test splits. On each fold, $\frac{4}{5}^{ths}$ of the data form the training set, which we use to fit the Lasso-constrained, linear model coefficients, and the remaining $\frac{1}{5}^{th}$ is used to compute the test MSE. To compute the test MSE, we apply the learned coefficients to the held-out fraction of the data to generate our predicted $\overline{y}_i$. We additionally compute the train MSE (mean squared error on the training set) for comparison.

We iterated through each kth fold and each alpha (values between 10^{-4} and 10^{4}, with 20 samples equally spaced on a $\log_{10}$ scale) to calculate the Mean Squared Error (MSE) of the testing data and the predicted model fit on the training data. The best alpha was chosen as the one with the lowest (or second lowest, see ***Results***) test MSE score.

3 RESULTS

Our goal was to identify genes whose expression (1) was uniquely related to hippocampus *or* striatum, and (2) varied systematically with age within their respective regions. To achieve this, we first selected a subset of genes that exhibited differential expression in hippocampus and striatum within at least one out of the four available age groups (1, 6, 16, and 24 months of age, **Figure 1, Table 1**; see ***Methods: Data Preprocessing*** for our operationalization of "differential expression"). In each age group, an overlapping but distinct group of genes were differentially expressed according to our criterion across hippocampus and striatum. **Table 1** lists the 27 genes that were differentially expressed across these regions in each age group.

We then constructed linear models that regressed age onto either gene expression in hippocampus, or onto gene expression in striatum. Each model thus fit coefficients to 27 unique regressors. With each regression, we aimed to obtain a sparse subset of genes whose expression was strongly related to age. To this end, we subjected each linear regression to a Lasso (or L1-norm) penalty, where a constrained subset of genes should emerge with non-zero coefficients (Tibshirani, 1996) given an optimized regularization strength. To optimize this hyperparameter, we fit 20 separate regressions for each region (40 models total), with each regression sporting a different regularization strength, α (for a list of candidate α, see ***Methods: Lasso-Regularized Linear Regression***). We compute the held-out, test mean squared error (MSE) for each regression, average the test MSE across folds, and plot these values against the model's regularization strength α (**Figure 2**).

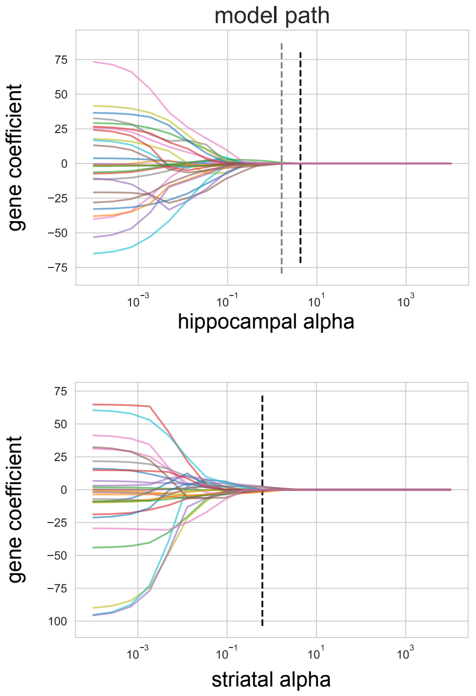

Figure 3: *Model path for each Lasso regression.* *In each panel, colored lines correspond to the value of a given gene's assigned coefficient as a function of regularization strength (alpha). Vertical black dotted lines mark the alpha that minimized test set mean squared error. For hippocampus, optimal alpha produces zero-valued coefficients for all 27 genes. Vertical grey dotted line marks the alpha immediately preceding optimal hippocampal alpha, which produces exactly one non-zero coefficient.*

The hippocampal model's optimal regularization strength (α = 4.28) produced zero-valued coefficients for all 27 genes (**Figure 3, *top***). To assess which hippocampal gene(s) would have incurred a non-zero coefficient when using a slightly less stringent regularization strength, we examined the coefficient distribution produced by the α that was one step size weaker (α = 1.62) and had only a marginally larger MSE relative to the optimal α (**Figure 4**, ***top***). In contrast to the hippocampal model, the striatal model's optimized regularization strength (α = 0.62) yielded several non-zero coefficients (**Figure 3, *bottom*, Figure 4 *bottom***).

A list of genes that were assigned non-zero coefficients is available in **Table 2**, with the corresponding hippocampal and striatal mean expression values across age and mice, and their assigned coefficient value. Notably, each gene had a negative ($\log_2$-transformed, z-scored) mean expression value, suggesting that these genes have generally low expression values relative to the full set of 8, 932 genes, and some of these selected genes (8430437G11Rik, and Jph1) *increase* their expression over the course of aging, while others (2610509G12Rik) *decrease* their expression over the course of aging.

4 DISCUSSION

This work leveraged methods in dimensionality reduction in order to parse out the most "relevant" genes, where relevance corresponds to systematic variation in gene expression with age. However, through our efforts in data exploration, we recognize the impact of pre-selection methods on the outcomes of the analysis. First, we chose to analyze genes from two specific brain regions, the hippocampus and striatum, and our pre-selection criteria specifically relied on differential expression across these two regions. We could have, alternatively, selected genes to analyze according to their descriptions within the three gene ontogeny domains included with the AGEMAP dataset (CC: cellular component, BP: biological process, and MF: molecular function, **Table 3**). Had we pursued this strategy, we may have explicitly selected genes related to synaptic plasticity and ion channels that predominantly appear in either region and are implicated in age-related cognitive decline, thus testing our original hypotheses more directly. Following yet another strategy, we could have used a more agnostic and exploratory approach to find the most age-relevant genes by fitting a unique model for each gene, where a given gene's expression level is regressed against age, as others have done before (Zahn et al., 2007; Van et al., 2021). While this approach may be successful when large amounts of observations are available per gene, per age group, the limited nature of the AGEMAP dataset (10 observations for a given gene, per age group) might be susceptible to producing noisy estimates, and the large number of tests performed (as many as there are genes, 8, 932) might have incurred a large probability of false negatives. Previous work in this vein has documented these difficulties and authors have consequently warned of the caveats with this strategy when analyzing this particular dataset (Van et al., 2021).

Although our particular modeling strategy is similarly susceptible to noisy coefficient estimates due to limited data, we constrain our search for relevant genes substantially by pre-selecting a subset of genes, and we apply rigorous (cross-validated) regularization strengths to increase the likelihood that only the most reliably related genes are assigned non-zero coefficients. These procedures reduce the probability of identifying spurious relationships in the data.

Hippocampal coordination with distributed regions of cortex is necessary for the formation and retrieval of memories. Similarly, the striatum plays a role in a diverse set of cognitive processes, including motor planning and action. Cell dysfunction in these regions is associated with severe disorders of memory (Alzheimer's Disease), movement (Parkinson's Disease), and cognitive flexibility (Anacker

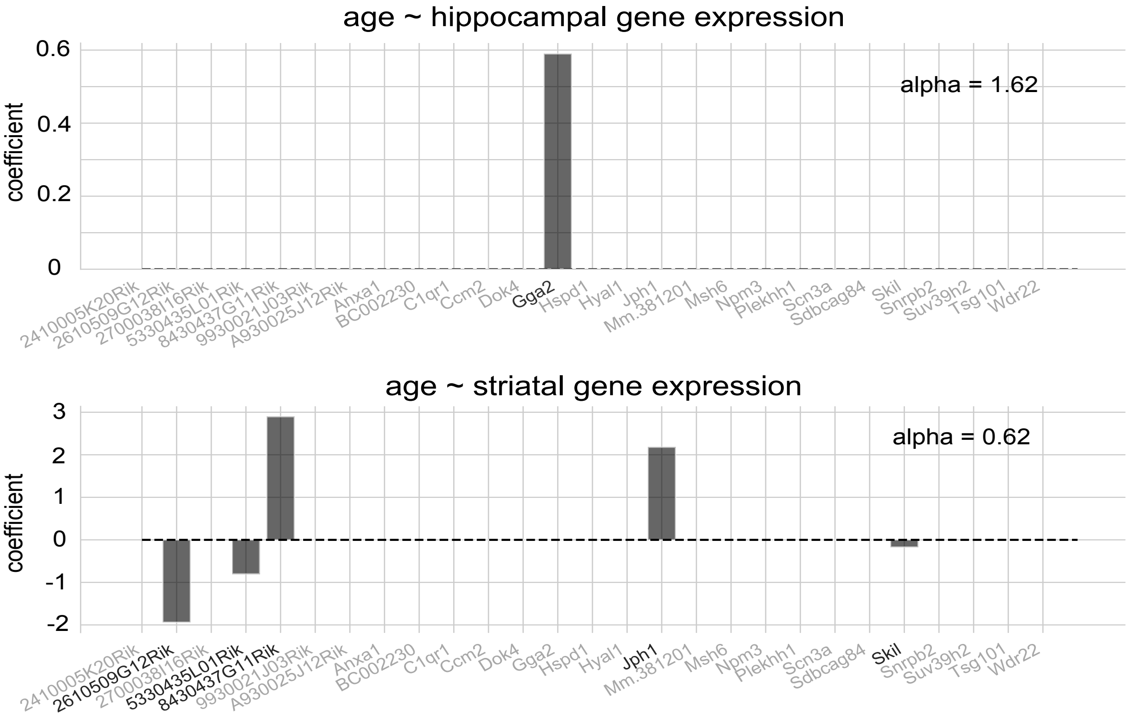

Figure 4: ***Coefficients for hippocampal and striatal models fit with near-optimal and optimal regularization strengths (alpha), respectively.*** *(top) Age regressed against hippocampal gene expression (27 genes), for model fit on entire dataset, with the near-optimal alpha. Coefficients for the optimal alpha are all zero. Non-zero coefficient for gene:* ***Gga2****. (bottom) Same as above, but age regressed against striatal gene expression, and fit with optimal alpha. Non-zero coefficients for genes:* ***2610509G12Rik, 5330435L011Rik, 8430437G11Rik, Jph1, Skil****.*

et al. 2017). In considering the genes that might be implicated in age-related cognitive decline, previous research finds that dysfunction of *cell circuits* appear to be more directly contributing to cognitive decline than cell death alone (Lee et al. 2010). Hippocampal and striatal gene expression associated with cell functions that are attenuated in old age are those involved in dopaminergic interactions and DNA damage repair (Bäckman et al. 2000, Xu et al. 2007), as well as cellular inflammation and immune response (Blalock et al. 2003). In the present study, the genes that emerged as relevant to the aging mouse brain seem to be largely involved in cellular processes related to cell growth, energy, and maintenance. Further research might investigate the role of these genes in cell processes in the hippocampus and the striatum, specifically, in order to elucidate their mechanistic role in behavioral deficits associated with old age.

5 REFERENCES

Anacker, Christoph, and René Hen. "Adult hippocampal neurogenesis and cognitive flexibility—linking memory and mood." Nature Reviews Neuroscience 18.6 (2017): 335-346.

Bäckman, Lars, et al. "Age-related cognitive deficits mediated by changes in the striatal dopamine system." *American Journal of Psychiatry* 157.4 (2000): 635-637.

Bettio, Luis EB, Luckshi Rajendran, and Joana Gil-Mohapel. "The effects of aging in the hippocampus and cognitive decline." *Neuroscience & Biobehavioral Reviews* 79 (2017): 66-86.

Blalock, Eric M., et al. "Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment." *Journal of Neuroscience* 23.9 (2003): 3807-3819.

de Jong, Laura W., et al. "Ventral striatal volume is associated with cognitive decline in older people: a population based MR-study." *Neurobiology of aging* 33.2 (2012): 424-e1.

Foster, Thomas C. "Involvement of hippocampal synaptic plasticity in age-related memory decline." *Brain Research Reviews.* (1999): 236-249.

Lee, Cheol-Koo, et al. "Gene-expression profile of the ageing brain in mice." *Nature genetics* 25.3 (2000): 294-297.

Machulda, Mary M., et al. "Practice effects and longitudinal cognitive change in normal aging vs. incident mild cognitive impairment and dementia in the Mayo Clinic Study of Aging." *The Clinical Neuropsychologist.* (2013): 1247–1264.

McCarthy, M., et al. (2011). Striatal origin of the pathologic beta oscillations in Parkinson's disease. Proceedings of the National Academy of Sciences, 108(28), 11620-11625.

Shen, W., et al. (2005). "Cholinergic suppression of KCNQ channel currents enhances excitability of striatal medium spiny neurons." *Journal of Neuroscience*, 25(32), 7449-7458.

Table 2: *Summary of model results. Note, mean is taken across ages and mice, for* log_2 *and z-scored expression values.*

gene name	hippo mean expression / coefficient	striatum mean expression / coefficient
Gga2	-1.31 / 0.59	-0.72 / 0
2610509G12Rik	-1.95 / 0	-2.88 / -1.94
5330435L01Rik	-3.15 / 0	-1.52 / -0.80
8430437G11Rik	-3.84 / 0	-2.99 / 2.90
Jph1	-2.24 / 0	-2.77 / 2.18
Skil	-3.82 / 0	-2.71 / -0.17

Table 3: *Summary of corresponding gene ontologies from results. MF = molecular function; BP = biological process; CC = cellular component.*

Gene Name	Function(s)
Gga2	**MF:** protein binding **BP:**Golgi to plasma membrane protein transport **CC:** Golgi apparatus
2610509G12Rik	**MF:** ATP binding, ATPase activity **BP:** chromatin organization **CC:** nucleoplasm; nucleus
5330435L01Rik	**MF:** chemotaxis **BP:** involved in synapse assembly **CC:** synapse, plasma membrane
8430437G11Rik	**MF:** enables molecular function **BP:** – **CC:** integral component of membrane
Jph1	**MF:** mediate cross talk between cell surface and intracellular ion channels **BP:** acts upstream of or within muscle organ development **CC:** junctional sarcoplasmic reticulum membrane
Skil	**MF:** cell growth and differentiation **BP:** chromatin and protein binding **CC:** nucleus

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.

Van, Monica, et al. "Changes in hippocampal gene expression of mice as a function of age" COGS 260 Mini-Project 1, UCSD. (2021)

Xu, Xiangru, et al. "Gene expression atlas of the mouse central nervous system: impact and interactions of age, energy intake and gender." *Genome biology* 8.11 (2007): 1-17.

Zahn, Jacob M, et al. "AGEMAP: a gene expression database for aging in mice." *PLoS Genet.* (2007)