Cogs 260, Spring 2021: Mini-project #1
Due date: May 2nd, 2021
Extended due date: May 6, 2021
First author: Monica Van
Authors: Holly(Yueying) Dong, Davis Lee, & Quirine van Engen

Changes in hippocampal gene expression of mice as a function of age

1. Introduction

Memory, an integral part of cognition, is known to decline with age. It is thought that the CNS transcriptome responds to aging in a region-specific manner [1]. The hippocampus, a brain region highly involved in the formation of new memories, warrants further investigation to better understand the aspects of memory that changes with age. A targeted examination of differentially expressed genes (DEGs) may reveal potential mechanisms of memory and cognition in regards to ageing.

As biological analogs for humans, the common house mouse, *mus musculus*, serves as an ideal candidate for our purposes. With well characterized life phase equivalencies, mouse models can provide suitable genetic analogs to the human genome in relation to various phases of lifespan in a condensed timeline. We analyzed the AGEMAP dataset published by Zahn et al. (Plos Genetics, 2003) [5] which utilized a microarray based cDNA detection to examine gene expression across four age groups of mice (1-month, 6-month, 16-month, 24-month) equivalent to Young, Mature Adult, Middle-aged, and Old respectively.

In the interest of exploring age-related gene expression changes and its connection to memory formation capabilities of the hippocampus, we performed an initial examination of DEGs over a mouse's lifespan by fitting a statistical regression model over the dataset. The top ten genes most correlated with age identified by our model were then further investigated using Gene Ontology analysis for their biological and molecular functions.

Our analysis compared multiple linear regression models based on age and sex parameters to examine predicted changes in gene expression. While the age only model had the highest MSE score, it only edged out other models by a small margin. Identification of top ten DEGs and subsequent GO analysis did not reveal any significant pathway level changes, but demonstrated some trends in early molecule development associated with metabolism and chromatin accessibility/ remodeling.

2. Methods

2.1 Datasets

For this project, we used the hippocampus dataset from AGEMAP [5]. This contains gene expression levels recorded as $\log_2$ transformation of their Z-scores from various tissues of mice at different ages (1-month, 6-month, 16-month, 24-month), as well as from both female and male mice. All mice were reported to have been fed ad libitum.

The hippocampus dataset contains 10 samples for each age category (five females and 5 males, totaling a sample size of N = 40. The microarray interrogated 8936 genes, including some polymorphisms.

We utilized various gene ontology databases (geneontology.org), genome sequencing databases (uniprot.org / genecards.org), and published research articles to examine genes of interest. Table 3 documents the top 10 identified genes loosely associated with age and their general functions as described by prior databases.

2.2 Pre-processing the data

Preparation of the AGEMAP dataset was pre-processed using the following steps:

1. Transpose the csv table into long format, such that each mouse sample comprises the rows and each the mice's age, sex, and expression per gene comprise the columns
2. Drop columns that lack an associated gene name
3. Rename columns with duplicate names to reflect different gene isotopes
4. Add the prefix, 'X,' to all gene names except Age and Sex to avoid Python errors associated with column names that start with numerics
5. Remove special characters from gene names
6. Convert the dataframe string values into floats and replace 'Sex' with dummy variables: Female becomes 0; Male becomes 1

2.3 Model description + protocol

Model 1: "Changes in gene expression as a function of age"

- Simple linear regression over age for each gene, pooled over sex:

$$Expression_{gene_i} \sim \beta_0 + \beta_{age} * x_{age} + \varepsilon$$

Model 2: "Changes in gene expression as a function of sex"

- Multiple linear regression over age and sex for each gene:

$$Expression_{gene_i} \sim \beta_0 + \beta_{sex} * x_{sex} + \varepsilon$$

Model 3: "Changes in gene expression as a function of age and sex"

- Multiple linear regression over age and sex for each gene:

$$Expression_{gene_i} \sim \beta_0 + \beta_{age} * x_{age} + \beta_{sex} * x_{sex} + \varepsilon$$

Model 4: "Changes in gene expression as a function of interaction between age and sex"

- Multiple linear regression over age and sex for each gene:

$$Expression_{gene_i} \sim \beta_0 + \beta_{age} * x_{age} + \beta_{sex} * x_{sex} + \beta_{interax} x_{age} x_{sex} + \varepsilon$$

Model 5: "Changes in gene expression as a function of natural-log-transformed age"

- Simple linear regression over natural-log-transformed age for each gene:

$$Expression_{gene_i} \sim \beta_0 + \beta_{LogAge} * ln(x_{age}) + \varepsilon$$

The models provide a coefficient for each gene and parameter (slope of gene expression over age) and a corresponding intercept. We then compare models to identify the model with the lowest average error for predicting expression levels across the genes. The next step was to verify if the genes with the lowest error on the best model have been identified to be important for aging.

2.4 Model comparison

For our cross-validation, we used the leave one out cross validation (LOOCV) method in consideration of the small size of our dataset (N=40).

For each gene, we took into account the four different models outlined in section 2.3. Each model utilized a different subset of the predictor variables (i.e. 'Age', 'Sex', 'Age+Sex', 'Age*Sex', 'Log(age)'). The different subsets were used alongside with their respective target values to fit a linear regression model. For each LOOCV iteration per model, we returned the Mean Squared Error (MSE) between the testing data point and values predicted by the model fit on the training data set, storing the result of all the

gene-model MSEs. The average MSE over all genes was taken for each model to compare which model performed best on the dataset. The model with the lowest average MSE was then chosen for further investigation.

2.5 Software

All code was executed on Python. Data cleaning made use of the Pandas, Numpy, and Regular Expressions packages. Data visualization used the Matplotlib and Seaborn libraries. Exploratory modelling leveraged Statsmodels, while actual model training, selection, and application employed the Scikit-Learn library. Other packages to assist in code execution include tqdm and sys.

3. Results

3.1 Model comparison

Table 1: The average MSE scores listed per model. The age model returns the lowest average MSE.

Model	Average MSE
Age	0.02323476273557134
Sex	0.02324919542693951
Age + Sex	0.024053112534480718
Age * Sex	0.024053112534480718
Ln(Age)	0.023257469780907904

Output from model comparison are error rates from test and training datasets. MSE was log transformed for an easier visual comparison. The MSE for the age model had the lowest MSE, but all MSE's were in the same range and distribution (Figure 1).

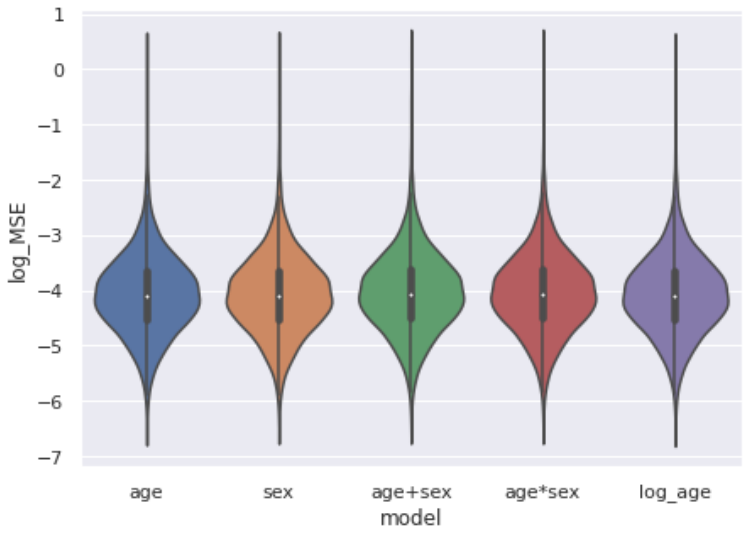

Figure 1: Model performance comparison. Violin plots represent the different models and their Mean Squared Error (MSE) log transformed. Log_age is the natural log of age. There are no significant differences between regression models, the age model edged out all other models by a small margin.

Table 2: Top 10 genes and their Linear Functions. Using the age model, which had the lowest average MSE, we returned the linear functions of the top 10 genes:

Gene	Linear Function
Pgm2l1	**0.43605858 - 0.00128745 x_{age}**
Inadl	**0.51468786 + 0.0004234 x_{age}**
Atrx	**-0.36813365 - 0.00138638 x_{age}**
Orc4l	**-0.02788113 + 0.00039024 x_{age}**
Itgb4bp	**-0.04581374 - 0.0005722 x_{age}**
AA407526	**0.29101507 + 0.00022538 x_{age}**
Kit	**0.23355112 - 0.00047681 x_{age}**
Tlk2	**-0.2951967 - 0.0009575 x_{age}**
Foxp1	**-0.54966215 + 0.00026706 x_{age}**
H19	**0.183593 + 0.00248613 x_{age}**

3.2 Visualization

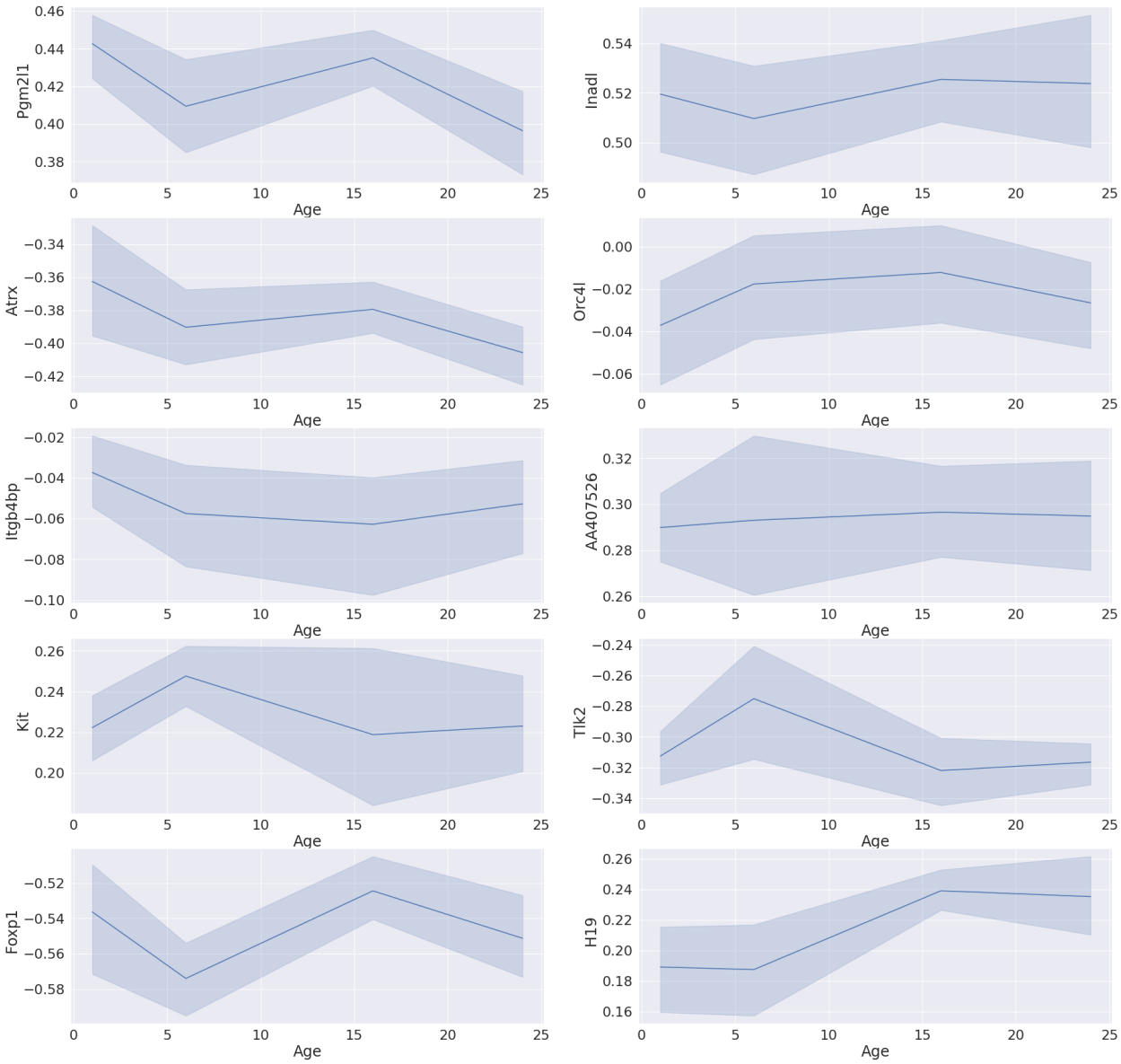

Figure 2: Gene expression over age for the top 10 genes found from our regression model including only age as predictor. Each age category (1, 6, 12, 24 months) has 10 data points. The gene expression quantified as log transformed, normalized z-scores as outlined in Cheadle et al 2003. While our model is linear, the data of these genes follow a wave like trend. Error bars in transparent blue indicate 95% confidence interval.

3.3 Gene ontology

Table 3: The following 10 genes were identified with the lowest MSE. The gene names, MSE, trend, and gene ontology are represented.

Gene name	MSE	Trend	Gene Ontology
Pgm2l1	0.001408	High in young and middle aged, dips in mature adult ages and decreases in old ages	Main biological functions involved in glucose and galactose metabolic processes. Molecular functions include intramolecular transferase activity, phosphotransferases, and glucose-1,6-bisphosphate synthase activity. Important paralog of this gene is PGM2 which has similar metabolic processes and phosphotransferase activity
Inadl	0.001504	Slight trend upwards with age	Also known as PATJ. Implicated in the production of a scaffolding protein that facilitates the localization of proteins to the cell membrane. Critical for the formation of tight junctions and epithelial apico-basal polarity.
Atrx	0.001631	Down regulated with age	Involved in transcription regulation and chromatin remodeling (chromatin binding and helicase activity). Facilitates DNA replication and is required for efficient replication of a subset of genomic loci. Important paralog of this gene is RAD54L2 which produces DNA helicases that modulates androgen receptor dependent transactivation in a promoter-dependent manner
Orc4l	0.001634	Downregulated expression in young and old age while	Recognizes and binds to H3 and H4 trimethylation marks (H3K9me3, H3K27me3, H4K20me3), plays a critical role in the assembly of pre-replication complexes

		upregulated in mature and middle aged adults	necessary to initiate RNA replication
Itgb4bp	0.001651	Upregulated expression in young and old age while downregulated in mature and middle aged adults	Primarily involved in stimulatory translation initiation factors downstream of insulin/growth factors. In tissues responsive to insulin, controls fatty acid synthesis and glycolysis by exerting translational control of adipogenic transcription factors. Involved in ribosome biogenesis, exporting ribosomal subunits from the nucleus. Required for ROS-dependent megakaryocyte maturation and platelet formation: controls the expression of mitochondrial respiratory chain genes involved in ROS synthesis.
AA407526	0.001692	Flat expression level trends throughout aging	Also known as Guf1, which is a GTPase. Functions as a secondary messenger system for various cellular functions.
Kit	0.00189	Upregulated in mature adult	Cell-surface receptor that plays an important role in the regulation of cell survival and proliferation, hematopoiesis, stem cell maintenance, gametogenesis, mast cell development, migration and function, and in melanogenesis.
Tlk2	0.001890	Upregulated in mature adult	Encodes a nuclear serine/ threonine kinase thought to function in the regulation of chromatin assembly by regulating the levels of histone H3 and H4. Associated with double-strand break repair of DNA damage. Highly expressed in embryos throughout development. Important paralog of this gene is TLK1 which has similar functions in encoding serine/

			threonine kinase involved in chromatin assembly. GO annotations related to TLK1 include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity.
Foxp1	0.001908	High in young and middle aged adult, dips in mature adult ages and decreases in old ages	Transcriptional repressor (binds to DNA and inhibits gene expression).
H19	0.001924	Steady upregulation with age	Maternally imprinted gene that may initiate pyroptosis and attenuates apoptosis.

4. Discussion

Summary of results

From our simple linear regression model,

$$Expression_{gene_i} \sim \beta_0 + \beta_{age} * x_{age} + \varepsilon,$$

we identified the top 10 genes that are most closely correlated with age. Overall, though there exists a weak trend between age and the genes identified, the sparseness of the AGEMAP dataset negatively impacts the power of the model.

Moreover, all other models have almost the same level of predictive power, with the interaction of Age and Sex in the Age * Sex model having a slighly higher average MSE. We interpret this as on average, the interaction between Age and Sex do not more accurately predict gene expression compared to sex or age (or log-transformed age) alone (Table 1).

Interestingly, 3 out of 10 top genes identified were associated with metabolism, chromatin accessibility, and methylation modification. DNA methylation is commonly thought to control the accessibility of genes, leading to varying genetic transcription levels. Prior research has established that DNA- methylation states change throughout ageing in a consistent and predictable manner such that methylation states have been proposed as biomarkers for ageing and age-related diseases. Such changes in methylation states in relation to ageing have led to altered gene expression levels of various genes throughout the lifespan of an individual. Increased methylation decreases chromatin accessibility and consequently gene expression, the reversal is true as well, where decreased methylation leads to increased gene expression.

Further analysis of hippocampal specific methylation changes through age is needed to elucidate whether there is a clear correlation between the two.

Theoretical and methodological explanation

The average MSEs of all the models are very similar, none of the models performed significantly better or worse. One explanation is the approach we took, which averaged over all the genes in the dataset, may have potentially reduced any significant correlative effects of potential genes. Our linear model may have missed such genes whose expression levels were particularly well-predicted, but were drowned out by the other ~8,000 gene-specific performances.

Similar to Zahn et al's original 2007 paper, we could not draw any significant correlative relationships between ageing and gene expression. This may be partly due to

the fact that the AGEMAP dataset had a maximum of 2 predictors for a total of 40 sparse data points, which did not provide enough power for our statistical models to draw any significant conclusions. Based on the trends observed in the top ten genes, the true relationship between ageing and gene expression may be more complicated or nonlinear in nature (Figure 2). In order to further probe the relationship between ageing and hippocampal gene expression changes, more data points are needed.

What we have learned

During this project, we have learned and implemented regression analysis in Python in working with a gene expression dataset. Creating and fitting a model for each gene and then averaging over all genes is not the best practice to tease out an optimal model. On the other hand, enormous gene datasets are difficult to work with, each gene may require its own most optimal model to capture it's trajectory. In future studies, it would be better to select a few genes of interest to constrain ourselves prior to model selection. However, doing so would lose the data-driven initial steps that can help in exploring obscure genes that may not have been associated with certain phenotypes.

Our efforts also revealed tradeoffs involved with the certain coding packages regarding Python. While the Linear Regression package from the SKLEARN library is convenient in building a model and evaluating the pipeline using Leave One Out cross-validation, it was not as statistically informative as the Ordinary Least Squares (OLS) methodology from the Statsmodels library. SKLEARN, unlike OLS which provides a quick summary including the p-values and confidence intervals, only provides the intercepts and coefficients of the fit model. The streamlining that SKLEARN provided required much less manual coding overall, but the library's packaging of several steps into a single command line also made methods to evaluate our models much more obscure to understand.

Impact of our results

The results documented here cannot claim any big impact on the ageing and genomics fields. However, the basic methodologies of our analysis could prove fruitful, as data-driven approaches without constraining analysis to a subset of genes may uncover previously overlooked factors. This approach may uncover genes that may not be directly related, but have an indirect effect from aging.

Future Steps

In the future, we could implement principle components analysis (PCA) before model selection in order to reduce the number of genes that we factor into our averaged performance score per model. Additionally, we could look at specific age ranges related to various stages of biological development instead of incorporating the whole age range. Genes associated with early development might not necessarily play a negative role in adult aging processes which was not accounted for in our current analysis. Moreover, it would be worth investigating whether hippocampal methylation level changes through ageing, given that 3 out of 10 of the identified genes in this project were associated with chromatin accessibility and methylation.

Another approach to the AGEMAP dataset could also switch the feature and target variables. In the project detailed here, we attempted to predict gene expression based on age, sex, their combination, or interaction. Alternatively our analysis could have configured our dependent variable to be age and used combinations of the ~8,000 genes as independent variables in order to see which would best predict biological age as a study of biomarkers of ageing.

Closing statement

Exploratory driven data-analysis on an entire genome could reveal previously unidentified genes associated with ageing. Such bottom-up methodologies may further drive unique hypotheses for the field.

References

1. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 2021;49(D1):D325-D334

2. Salameh, Y., Bejaoui, Y., & El Hajj, N. (2020). DNA Methylation Biomarkers in Aging and Age-Related Diseases. Frontiers in Genetics, 11. doi:10.3389/fgene.2020.00171

3. Johnson AA, Akman K, Calimport SR, Wuttke D, Stolzing A, de Magalhães JP. The role of DNA methylation in aging, rejuvenation, and age-related disease. Rejuvenation Res. 2012;15(5):483-494. doi:10.1089/rej.2012.1324

4. Xu, X., Zhan, M., Duan, W. et al. Gene expression atlas of the mouse central nervous system: impact and interactions of age, energy intake and gender. Genome Biol 8, R234 (2007). https://doi.org/10.1186/gb-2007-8-11-r234

5. Zahn JM, Poosala S, Owen AB, Ingram DK, Lustig A, Carter A, Weeraratna AT, Taub DD, Gorospe M, Mazan-Mamczarz K, Lakatta EG, Boheler KR, Xu X, Mattson MP, Falco G, Ko MS, Schlessinger D, Firman J, Kummerfeld SK, Wood WH 3rd, Zonderman AB, Kim SK, Becker KG. AGEMAP: a gene expression database for aging in mice. PLoS Genet. 2007 Nov;3(11):e201. doi: 10.1371/journal.pgen.0030201. Epub 2007 Oct 2. PMID: 18081424; PMCID: PMC2098796.

6. Richardson, B. (2003) Impact of aging on DNA methylation. *Ageing Research Reviews*. https://doi.org/10.1016/S1568-1637(03)00010-2

7. Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of Microarray Data Using Z Score Transformation. The Journal of Molecular Diagnostics, 5(2), 73–81. doi:10.1016/s1525-1578(10)60455-2

8. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9