Understanding the genetic basis of cognitive reserve towards an Alzheimer's cure

Project SBI159

Introduction

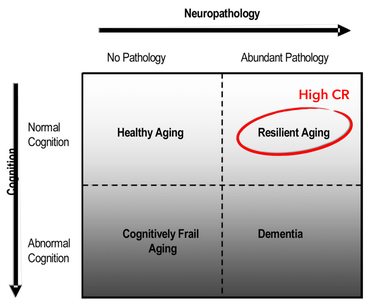
- **Alzheimer's disease** (AD) is a devastating disease that affects over 5 million Americans – and this number is only getting worse with increased life expectancy
- **Cognitive reserve** (CR) is a trait that represents natural resilience to diseases that cause neuropathological damage
  - Slows the onset of cognitive decline due to AD, dementia, and aging in general
  - Processes that might be associated with CR include
    - **Neurogenesis** – neuron creation by neural stem cells
    - **Neuroplasticity** – the ability of the brain to adapt & change
- Understanding the underlying mechanisms for CR could be a path to developing the "holy grail" – drugs to prevent and cure AD
  - Gain insight into how AD works, and why some people are able to resist it while others aren't
  - **If CR is genetic,** develop epigenetic treatments that can increase CR to prevent or reverse the onset of AD and other aging-related diseases

Hypothesis & Goals

- **I hypothesize that CR is genetic: variation in the expression of genes related to neuroplasticity and neurogenesis contributes to variation in cognitive reserve**
- Conduct a large-scale analysis of the relationship between CR and gene expressions, with two goals:
  - Find gene expression biomarkers that are predictive of CR, suggesting that these genes are responsible for CR and ought to be targeted by epigenetic AD treatments
  - Gain insight into cell types and pathways involved in the development of these genes
    - Which neural functions support CR? Can we enhance these functions to potentially treat AD?

Background

- In 1989, Katzman et al. found that aging women had a normal cognitive phenotype (i.e., no cognitive decline) *despite* advanced AD pathology found at death, **first suggesting the existence of CR**
- Since then, studies comparing actual cognitive performance to postmortem diagnosis have shown that **up to 30% of individuals who *should* have AD and dementia (based on pathology) don't show symptoms**, implying high CR

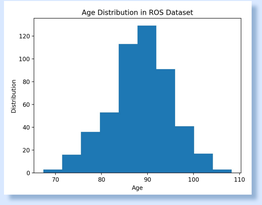


Source: *Negash, Selamawit, et al. "Cognition and neuropathology in aging: multidimensional perspectives from the Rush Religious Orders Study and Rush Memory and Aging Project." Current Alzheimer Research 8.4 (2011): 336-340.*

- Efforts to actually understand CR have been hindered by difficulties in defining, quantifying, and measuring it
  - Often associated with environmental factors such as educational attainment, IQ, and cognitive stimulation
  - But this isn't the full picture – although there's limited research associating CR with genetic factors, preliminary work has tied it to BDNF, a gene related to neuroplasticity
- My work uses a novel method to understand **whether** there's a genetic basis to CR, and, if so, **what specific genes are involved**
  - First time CR has been connected to large-scale gene expression data
  - I develop and use a novel method to quantify CR – a difficulty in past work

Dataset

- I used bulk tissue RNA-seq data from the **Religious Order Study**, a longitudinal study of aging and dementia
  - Participants had no known dementia at the time of study entry
  - Annual clinical evaluations recorded:
    - *Markers of AD pathology:* amyloid-β accumulation and neurofibrillary tangle accumulation
    - *Cognitive performance:* average of z-scores on 21 cognitive tests
  - 188 males, 314 females, ages 65 to 108
- After cleaning, 495 valid RNA-seq samples representing 19,983 genes
- Data was not divided into groups based on any covariates or demographic information
- Although information on factors such as educational attainment was included, the aim of this project was to analyze CR across the entire population without taking these factors into account.



Procedure, Methods & Results

1 Quantifying Cognitive Reserve

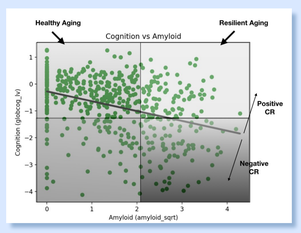
- I plotted cognition against **amyloid-β accumulation** and **neurofibrillary tangle accumulation**, both of which served as **proxies for AD**
  - I used a linear regression model of cognition for each AD proxy to predict cognition
  - **I quantified cognitive reserve by finding the difference between predicted and actual cognition for each sample**

Linear Regression

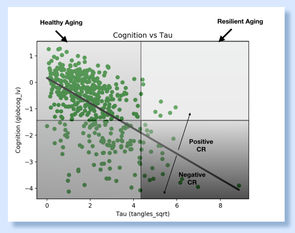
- I used a linear regression model to compute the difference between expected and observed cognition for a given severity level of AD
- I ran two different linear regressions
  - Predict cognition from the amyloid-β accumulation level (amyloid level) and from neurofibrillary tangle accumulation level (tau level)
  - **This is a novel way to quantify CR – a task that's caused difficulties in previous work**

I calculated cognitive reserve based on the predicted values for each linear regression run. For each sample, cognitive reserve was calculated as:

$y_{actual} - y_{pred} = CR$



r^2: 0.12551032694375386
Intercept: -0.27277342
Coefficient: -0.36668727
p-value - intercept: 0.000926
p-value - coeff: 4.39E-16



r^2: 0.215
Intercept: 1.24968
Coefficient: 0.553263
p-value - intercept: 1.39 E-37
p-value - coeff: 8.29E-28

2 Predicting Cognitive Reserve

I ran an **Elastic Net regression** on the data in order to create a predictor for CR based on gene expression. I computed three versions:

Predictor #1: Resilient Aging

- Positive CR was computed for all individuals with amyloid/tau greater than average (Abundant Pathology)

Predictor #2: Healthy & Resilient Aging - Positive CR values (Normal Cognition)

Predictor #3: All Aging - Both negative and positive CR values, all types of aging represented

Prediction Pipeline

A. Data Preparation

I discarded non-numerical values in variables of interest (amyloid and tangle buildup, global cognition) in the metadata

I randomly split the data into training (75%) and testing (25%) sets

I standardized gene expression values by removing the mean and scaling to unit variance

B. Feature Selection

I used feature selection to build a regression predictive model based on the high-dimensional data of 19,983 gene predictors and 495 samples

I used regularization to reduce overfitting and variance of the prediction error, and to handle correlated predictors. A combination of ridge regression and LASSO (least absolute shrinkage and selection operator) penalized models was utilized.

- LASSO represents the **ℓ1 penalty** that penalizes the sum of the absolute coefficients. LASSO leads to a sparse solution, selecting one gene from each group
- Ridge regression penalizes the **ℓ2 norm** of the model coefficients and keeps all the predictors in the model, shrinking them proportionally. Ridge regression keeps genesthat share the same biological pathway in the model.

C. Regression Model - Hyperparameter Selection

I used Elastic Net cross-validation on the training data to automatically select the best hyperparameters with iterative fitting along a regularization path to automatically tune the value of alpha

I used a list of l1 ratios where 0 < l1_ratio < 1 presents the penalty which is a combination of ℓ1 and ℓ2, leaving the alpha value to be automatically chosen by the Cross Validator. The different l1 ratios are tested by Elastic Net CV and the one giving the best prediction score is used.

D. Regression Model - Evaluation and Hyperparameter Tweaking

R-squared was computed on the trained model

Root Mean Square Error (RMSE) was computed on both the train and test data sets

To prevent overfitting, the input list of l1 ratios was tweaked to achieve as close a match as possible between the RMSE values on both train and test data sets, at the same time maximizing the R-squared on the trained model.

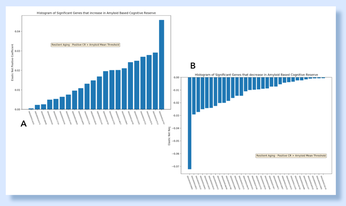
Observations

- In general, I observed that increasing the l1 ratio helped in reducing overfitting
- As the AD proxy data became more selective, the optimal l1 ratio increased

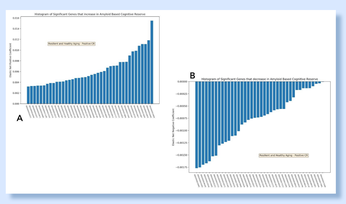
The most accurate predictor for CR was based on neurofibrillary tangle accumulation for Healthy & Resilient Aging

- Three distinct models of cognitive reserve prediction were created. Predictor #1 only took into account positive CR values from donors with severe AD (as determined by amyloid or tau level). Predictor #2 took positive CR values into account. Predictor #3 took all CR values into account.
- A list of genes that were predictive for cognitive reserve was generated for each model. The elastic net regression assigned a coefficient to multiply each gene's expression by to calculate CR. Those genes with the highest positive coefficients assigned to them would be predictive and potentially causal for high CR. Those with low negative coefficients would also be influential, but might work in the opposite direction

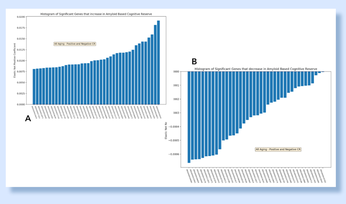
Amyloid



Predictor #1 - Resilient Aging
A consequence of the increased selectivity and high L1 penalty on this predictor led to fewer genes with nonzero coefficients. 54 genes were selected. 22 genes had a positive coefficient, while 32 had a negative coefficient.

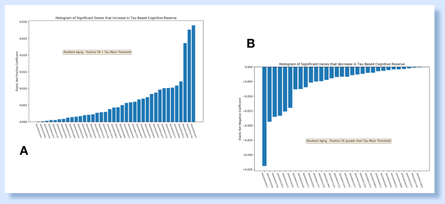


Predictor #2 - Resilient & Healthy Aging
196 genes with nonzero coefficients were found. 95 genes had a positive coefficient, while 101 had a negative coefficient.

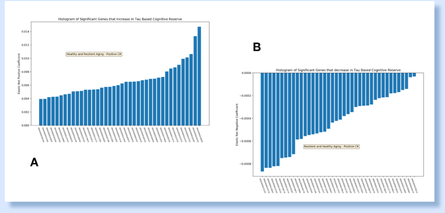


Predictor #3 - All Aging
537 genes with nonzero coefficients were found. 292 genes had a positive coefficient, while 245 genes had a negative coefficient.

Tau

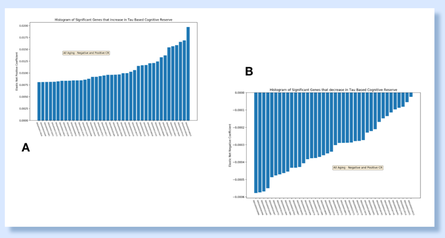


Predictor #1 - Resilient Aging
69 genes were selected. 38 genes had a positive coefficient, while 31 had a negative coefficient.



Predictor #2 - Resilient & Healthy Aging
258 genes with nonzero coefficients were found. 122 genes had a positive coefficient, while 136 had a negative coefficient.

r^2 (train): 0.525
RMSE (train): 0.315
RMSE (test): 0.363



Predictor #3 - All Aging
541 genes with nonzero coefficients were found. 281 genes had a positive coefficient, while 260 had a negative coefficient.

Procedure, Methods & Results

3 Finding Potential CR Biomarkers

Determining Biomarkers

- **A:** Genes with highest and lowest coefficients for amyloid and tau
- **B:** Genes common to all three predictors for amyloid and tau
- **C:** Genes common to Healthy & Resilient Aging across amyloid and tau

These potential biomarkers should be validated through further research.

	Biomarkers from highest coefficients (positive)	Biomarkers from lowest coefficients (negative)
Aeo/old - Resilient Aging	IGRT60	TCAF2P1
Aeo/old - Healthy + Resilient Aging	ANGPTL2	CCDC6A2
Aeo/old - All CR Values	ALDH1A1	HIF0
Tau - Resilient Aging		CD300E
Tau - Healthy + Resilient Aging	AL645608.2	C1DP1
Tau - All CR values	KTN1-AS1	UBE2F-SCLY

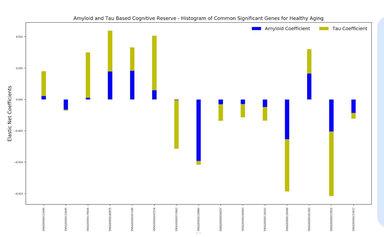
A

	Biomarkers from Predictor Overlap - Positive coefficient	Biomarkers from Predictor Overlap - Negative coefficient
Amyloid	CDT1 OR51I2 AL139352.1 ACTN1 ASGPTL2 RP11-1M121.3 SLC1A7	AC091132.4
Tau	LINC00960 KTN1-AS1 LINC01661	PITRM1-AS1

B

Biomarkers from Amyloid/Tau Overlap - Positive coefficient	Biomarkers from Amyloid/Tau Overlap - Negative coefficient
AC025048.2 TIMM8.1 SLC1A7 LINC00960 TEX29 MRGPRX3	IL17 160431.1 GSTM5 AC046185.1 AC008443.1 AC010270.1 IER3IP1 TCAF2P1

C



Common Genes

- **Coefficients of C Biomarkers:** Common genes from the amyloid and tau models for Resilient and Healthy Aging
- The common genes identified as biomarkers have the same direction: They increase or decrease for both proxies, further strengthening the correlation

4 Analyzing Potential Biomarkers

Pathway Analysis

- To determine the pathways associated with the potential CR biomarker genes, I performed a Gene Set Enrichment Analysis (GSEA)

Cell Type Marker Analysis

- I performed an overlap analysis to determine relevant cell types
- Cell type markers were determined by Habib et al. in a massively-parallel single-nuclei RNA-seq with DroNc-seq
- I converted the cell type marker lists from gene symbols to ensembl ID to find enrichment, and I used the Biological Database Net's (bioDBnet) Database to Database (db2db) conversion tool

5 Discussion

Common Genes

- These results suggest that CR has a genetic basis – common biomarkers found between predictors had coefficients that shared directionality

Pathway Analysis

- **GO terms across all predictors related to cell function.**
  - **This supports the idea that cognitive reserve is related to neurogenesis.** Increased neurogenesis may increase gene expression related to cellular functions in patients with high CR
  - Neurogenesis is well-known as a response to neurodegeneration. Studies have found that many treatment drugs for AD tend to increase neurogenesis in the brain. Finding a way to **extend neurogenesis** as degeneration worsens could be promising as a treatment.

Cell Type Marker Analysis

I analyzed the representation of each cell type in the data using Habib gene expression cell type markers. Neural stem cell (NSC) marker genes had a significant presence in the data, being found among genes that both increase and decrease as CR increases in the brain.**The representation of neural stem cells in the data supports the linkage of neurogenesis and neuroplasticity with CR.**

Conclusions & Summary

- **The findings of this project support the hypothesis that variation in gene expression related to neurogenesis and neuroplasticity contributes to cognitive reserve.**
  - Common genes were found between predictors whose coefficients shared directionality
  - **This suggests that CR has a genetic basis**
- **Pathway analysis suggests that neurogenesis and neuroplasticity are related to CR.**
  - Subjects with high CR might be able to maintain neurogenesis for a higher rate for a longer time, compensating neuronal loss experienced in AD
  - The cell type marker gene overlap analysis adds further credence to these results – **neural stem cells were highly represented in the data**
  - **Results suggest that neurogenesis and neuroplasticity are cell functions that support CR**
- **This project used novel methods to analyze CR**
  - A unique way of quantifying CR was developed
  - Models were built to predict CR from gene expression
  - Tau-based models were more accurate – neurofibrillary tangle accumulation seems to be a better AD proxy
  - **Potential gene expression biomarkers were identified for CR**
- **Genes that are causal for neurogenesis (and thus, high CR) might be potential treatment targets for future epigenomic drugs in order to target the root cause of AD**
  - Epigenomic therapies can regulate genes that increase CR, acting as a preventative treatment or even **a cure for AD**

Future Work

Future extensions of the project include applications in drug discovery, epigenomics, and treatments for AD. Determining the histone modifications responsible for affecting CR-enhancing genes opens up avenues for drug discovery research.

- Biomarkers should be validated through longitudinal studies on the mouse brain, and analyzed in the developing brain to gain insights on when AD develops in one's lifespan
- Further analysis on the links between AD and neurogenesis will help determine the relationship between CR and neurogenesis

I'd like to thank **Professor Andreas Pfenning, Easwaran Ramamurthy,** and the **Neurogenomics Lab at Carnegie Mellon University** for their advice with this project. I'm so grateful for their guidance.