Genome-Wide Association Study on Gene Pathway Identification and Cognitive Function Prediction

Li Xing[1], Kyle Gardiner[1], Mathew Zbitniff[1], Roham Asgari[1], Hanye Zhong[2]

1: University of Saskatchewan, 2: University of British Columbia

1. Introduction

Genome-Wide Association Studies (GWASs) are a research approach to identify genetic variations associated with complex diseases/traits of interest [1]. However, traditional GWAS methods lack power when sample sizes are small, leading to missed associations. Therefore, we employ machine learning (ML) methods to identify relevant single nucleotide polymorphisms (SNPs) and build a prediction model for outcome variables. To illustrate how to use ML for this role, we used a small dataset collected via the Alzheimer's Disease Neuroimaging Initiative (ADNI) to identify SNPs associated with Clinical Dementia Ratings (CDR). In addition, we performed gene-enrichment analysis to make interpretations of identified SNPs.

2. Challenges Faced

Performing high-dimensional inference with small sample sizes (n) and a large number of predictors (p), commonly stated $p >> n$ can produce inconclusive results. Our data, seen in the table below, only has 197 participants and approximately 640k SNPs.

	Normal Cognition (N=122)	Some Cogntive Impairment (N=75)	Total (N=197)
Age (years)			
Mean (SD)	69.0 (4.15)	69.5 (5.06)	69.2 (4.51)
Median [Min, Max]	67.9 [60.9, 82.2]	68.3 [61.6, 85.2]	67.9 [60.9, 85.2]
Ethnicity			
Not Hispanic/Latino	114 (93.4%)	63 (84.0%)	177 (89.8%)
Other	8 (6.6%)	12 (16.0%)	20 (10.2%)
APOE.e4			
Zero Alelles	88 (72.1%)	56 (74.7%)	144 (73.1%)
At Least One Alelle	34 (27.9%)	19 (25.3%)	53 (26.9%)
MMSE			
Mean (SD)	28.4 (1.63)	27.8 (1.83)	28.2 (1.73)
Median [Min, Max]	29.0 [21.0, 30.0]	28.0 [22.0, 30.0]	28.0 [21.0, 30.0]
ADAS			
Mean (SD)	10.3 (4.45)	13.7 (5.03)	11.6 (4.95)
Median [Min, Max]	10.0 [3.00, 23.0]	13.0 [4.00, 28.0]	11.0 [3.00, 28.0]

Due to the sample size, traditional single SNP-wise association testing resulted in no significant results, as depicted in the manhattan plot below. Our solution to these inconclusive results was to use various ML methods (laid out in the approach section) to identify relevant SNPs that may have been missed.

3. Approach

[**Step 1**]: Logistic regression was used to identify the SNPs most associated with CDR. The model was adjusted by covariates including age, ethnicity, ApoE4, and 2 other cognitive measures; Mini-Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale (ADAS)

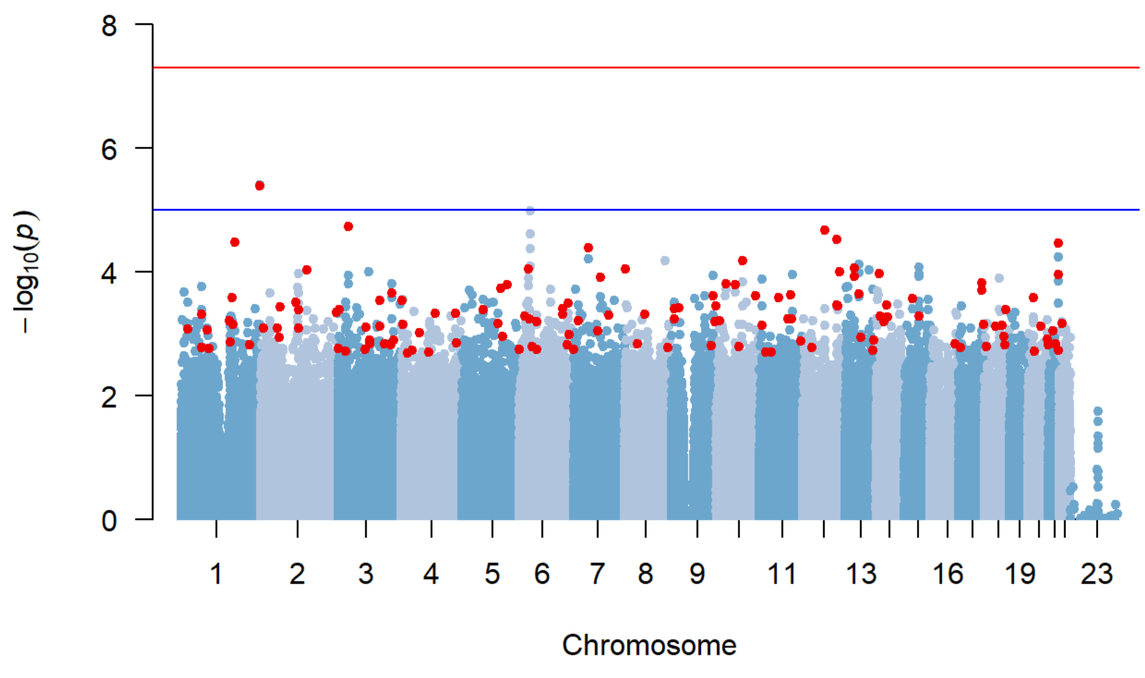
[**Step 2**]: Elastic-net penalized regression was used on the top 1000 SNPs from logistic regression to select important SNPs [3] to minimize:

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

[**Step 3**]: The elastic net selected SNPs would then be used in the following gene-enrichment mapping steps:

4. Results

As mentioned, traditional single SNP-wise association testing didn't produce any significant results. This manhattan plot shows none of the SNP points extend beyond the red horizontal line which represents the genome-wide significance threshold (p-value=$5 * 10^{-8}$). After using the top 1000 SNPs for elastic net, 136 were selected to have some association with CDR. The red dots on the plot represent the SNPs selected via elastic net.



Gene enrichment analysis was used on the 136 elastic net selected SNPs. First, the SNPs were mapped to their genes, then the genes were mapped to the biological pathways. The most prominent path identified was the positive regulation of interleukin-2 (IL-2). Studies have concluded lower levels of IL-2 are associated with declined cognitive scores. They even found that IL-2 levels may be better indicators of cognitive decline compared to the traditional Aβ and tau biomarkers [4].

5. Conclusions

Traditional single SNP-wise association tesing lacks power when dealing with small sample sizes. However, this report demonstrates the ability to utilize other approaches such as ML to identify significantly associated biomarkers. In addition, gene-enrichment analysis was utilized to find biological pathways associated with cognitive impairment/decline to aid in interpreting the selected SNPs. Potentially, this research will foster the use of other ML techniques in future GWAS when presented with the challenges of small sample sizes. Hopefully, upon realizing the potential, ML techniques will become an integral aspect of future GWASs.

6. Future Directions

1. Use a similar method for the other two cognitive measures to identify associated SNPs. Check for overlapping SNPs between measures.
2. Use other ML techniques capable of variable selection
3. Test to check if this method can be generalized to other datasets of varying sizes.

7. References

[1] Emil Uffelmann, et al. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), 2021.

[2] Shaun Purcell, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[4] Chih-Sung Liang, et al. Better identification of cognitive decline with interleukin-2 than with amyloid and tau protein biomarkers in amnestic mild cognitive impairment. *Frontiers in Aging Neuroscience*, 13, 2021