GWAS Analysis of Alzheimer's Disease: Identifying High-Impact Risk Loci

Xingling Wan, MSc Statistical Data Science, UoB

1) Project Summary & Objectives

Alzheimer's disease is a chronic, long-term neurodegenerative disorder with complex etiology, whose progression typically worsens over time, necessitating sustained, in-depth research by scientists. Using published GWAS summary statistics (Wightman et al., 2021), this project reproduced key association patterns in AD and highlighted the strongest signal region on chromosome 19 in Manhattan plots.

2) Data Characteristics

Data Source: Wightman et al. (2021), *Nature Genetics* (PMID: 34493870).

Sample size: 1,126,563 individuals

Data Scale: Published GWAS summary statistics encompassing approximately **12.68 million variants (SNPs)**.

3) Methods

Analysis was performed using **R (version 4.5.2)**, leveraging data.table for high-speed I/O of 12.6M rows and qqman for robust genomic visualization. Detailed environment specifications and package versions are documented in results/session_info.txt within the GitHub repository.

Memory Management: Used explicit object removal (rm) and garbage collection (gc) to reduce peak memory usage during processing 12.6M rows.

Data processing: Within the computational constraints of personal computers, a downsampling strategy was adopted for visualization only. For indexing and plotting, variants were labeled using coordinate-based IDs (CHR:BP).

Plotting Strategy:

Manhattan Plot: Used significance retention ($P < 0.01$) combined with 0.5% random sampling of non-significant sites, balancing efficiency with visual integrity.

Q-Q Plot: For computational convenience, a 1% uniform sample was used to approximate the Q–Q distribution for visualization.

4) Findings

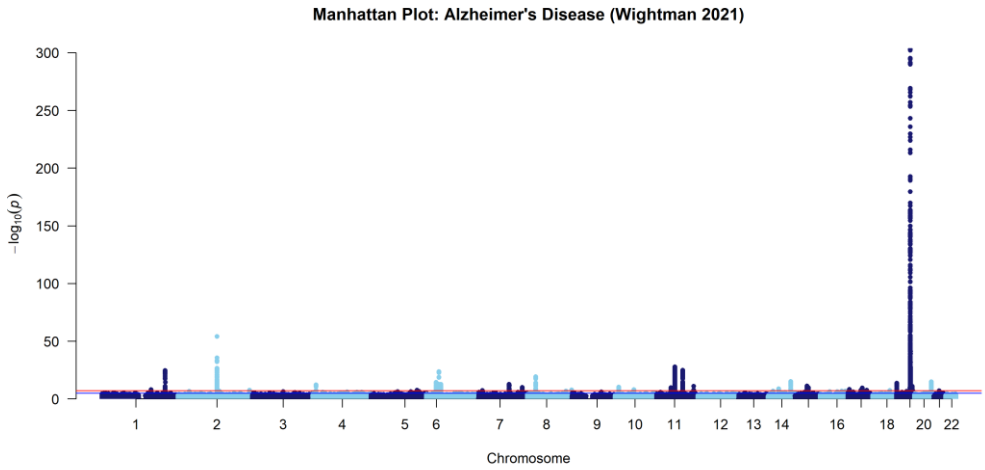

Figure1: Manhattan Plot

The Manhattan plot shows the strongest association signal on chromosome 19. This region is consistent with the well-known AD risk locus near APOE.

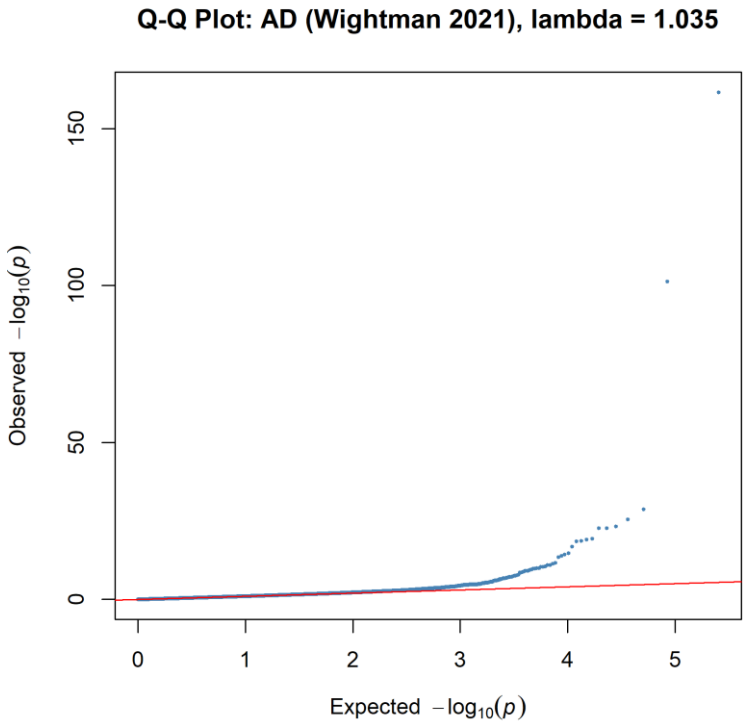

Figure2: Q-Q Plot

The genomic inflation factor was $\lambda_{GC} = 1.035$. Values close to 1 suggest minimal overall inflation at the genome-wide level.

5) **Conclusion & Core Competencies**

Scale: Successfully managed and analyzed 12M+ variants on a standard personal computer.

Integrity: Performed basic sanity checks on summary statistics and documented the computational environment (session_info.txt).

Interpretation: Interpreted diagnostic metrics (Q–Q/λGC) and highlighted key locus-level signals (chr19 region).

6) Limitations & Next Steps

This project focuses on summary-statistics level sanity checks and visualization; it does not include a full, standardised GWAS quality control (QC) pipeline. Additional systematic checks and locus-level follow-up would be needed for a full analysis.

Code Availability: The R code for this project is openly available at **GitHub: https://github.com/xingling1218/AD-GWAS-Analysis** , which *includes: scripts for* λ_{GC} *calculation and high-resolution Manhattan/Q-Q plots.*

Reference:

Wightman, D. P., Jansen, I. E., Savage, J. E., Shadrin, A. A., Bahrami, S., Holland, D., Rongve, A., Børte, S., Winsvold, B. S., Drange, O. K., Martinsen, A. E., Skogholt, A. H., Willer, C., Bråthen, G., Bosnes, I., Nielsen, J. B., Fritsche, L. G., Thomas, L. F., Pedersen, L. M., Gabrielsen, M. E., … Posthuma, D. (2021). A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature genetics*, *53*(9), 1276–1282. https://doi.org/10.1038/s41588-021-00921-z