

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. This project established a GWAS analysis workflow capable of handling traits of varying complexity. By applying the massively scaled Alzheimer's disease dataset developed by Wightman et al. (2021), it validated the conclusion that the most significant genetic risk locus in Alzheimer's disease is located on chromosome 19.

2) Data Characteristics

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

Sample size: 1,126,563 individuals

Data Scale: **Harmonized 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: Employed dynamic memory reclamation via `rm()` and `gc()` at any time, designed to prevent memory overflow and efficiently process datasets exceeding tens of millions of rows.

Data processing: To ensure efficacy and within the computational constraints of personal computers, a downsampling strategy was adopted, utilising coordinate synthesis (CHR:BP) to establish unique site-specific identifiers.

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: Utilised 1% global uniform sampling to ensure representativeness of the distribution.

4) Findings

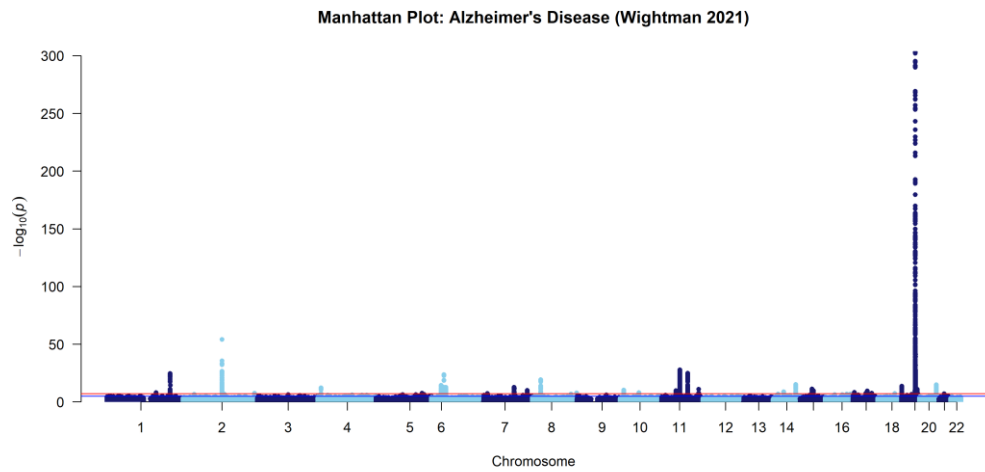


Figure1: Manhattan Plot

This study reveals a highly significant association signal for Alzheimer's disease located on chromosome 19. This peak corresponds to the APOE region in biological terms and represents the most important genetic risk locus in Alzheimer's disease.

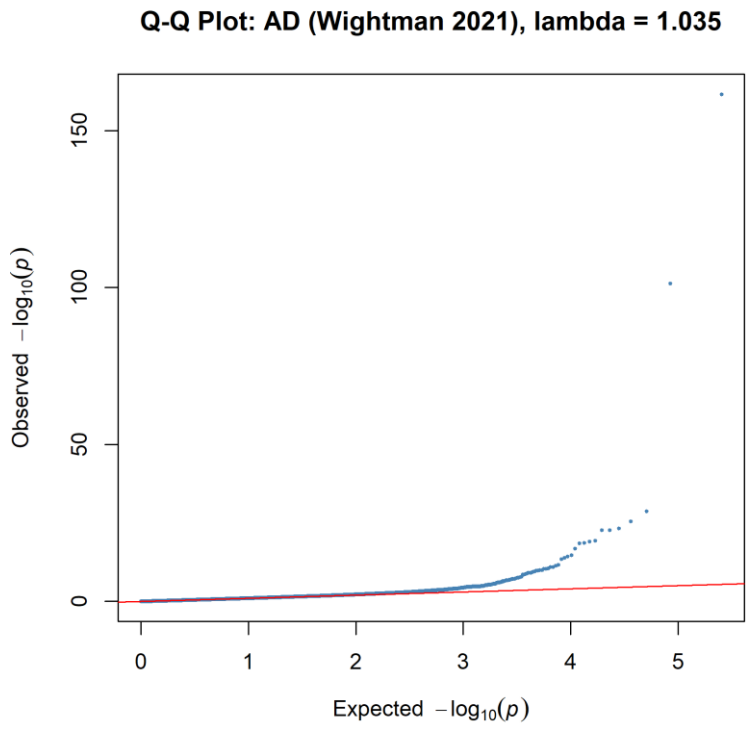


Figure2: Q-Q Plot

$$\lambda_{GC} = 1.035$$

Given that λ is very close to 1.0, this indicates that the study implemented exceptionally rigorous correction for population stratification when processing 1.1 million samples. This demonstrates that the observed association signals are robust and genuine biological signals, rather than statistical artefacts.

5) Conclusion & Core Competencies

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

Integrity: Demonstrated high data auditing skills by manually verifying dataset consistency across multiple biological databases.

Interpretation: Capable of bridging statistical metrics (λ) with biological insights (APOE).

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>