Non-linear modelling reveals novel epigenetic associations between DNA Methylation and Alzheimer's Pathology

Student Number: 720034352

Word Count: 5809

Table of Contents

Table Of Abbreviations

Abbreviation	Definition
AD	Alzheimer's Disease
BH	Benjamini-Hochberg
CV	Cross-Validation
DAVID	Database for Annotation, Visualisation, and Integrated Discovery
DNase	Deoxyribonuclease
EDF	Estimated Degrees of Freedom
EFS	Ensemble Feature Selection
EWAS	Epigenome-Wide Association Study
FDR	False Discovery Rate
fGSEA	Fast Pre-Ranked Gene Set Enrichment Analysis
GAM	Generalised Additive Model
GBM	Gradient Boosting Machine
GO	Gene Ontology
GREAT	Genomic Regions Enrichment of Annotations Tool
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MCL	Markov Clustering Algorithm
NES	Normalised Enrichment Score
NFTs	Neurofibrillary Tangles
NMDA	N-Methyl-D-Aspartate
PCA	Principal Component Analysis
PPI	Protein-Protein Interaction
R^2	Coefficient of Determination
REML	Restricted Maximum Likelihood
ROS/MAP	Religious Orders Study and Memory and Aging Project
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
λGC	Genomic Inflation Factor

Table Of Figures and Tables

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterised by cognitive decline and neuropathological changes, including amyloid plaques and neurofibrillary tangles. While epigenome-wide association studies (EWAS) have identified DNA methylation changes associated with AD, traditional linear models may overlook complex, non-linear relationships. This study applies Generalised Additive Models (GAMs) to model non-linear associations between DNA methylation and Braak staging at CpG sites. Analysis of prefrontal cortex tissue from the ROS/MAP cohort identified 534 significant CpG sites, 267 of which are novel. Homeobox genes were overrepresented among these loci, with several annotated to CpG sites which exhibited high feature importance, suggesting a substantial influence over Braak stage. Protein-protein interaction (PPI) analysis further revealed that some homeobox genes may influence multiple molecular networks. Functional enrichment analysis implicated these epigenetic modifications in neuronal function, synaptic regulation, and stem cell pluripotency. These findings demonstrate that non-linear models uncover novel epigenetic associations in AD, providing new insights for identifying CpG sites for biomarker discovery and therapeutic development.

Lay Abstract

Alzheimer's disease (AD) leads to memory loss and cognitive decline and affects millions globally. Changes in gene expression have been linked to AD and can influence brain function; however, previous studies have assumed these changes follow a linear pattern, potentially missing complex non-linear changes. This investigation searched for non-linear patterns between changes in DNA expression and Alzheimer's disease.

By analysing brain tissue from individuals who died with Alzheimer's, we identified 534 DNA sites linked to disease progression, including 267 new sites. Notably, a group of genes known as homeobox genes showed substantial changes. These genes were connected to multiple protein networks in the brain, including one extensive interconnected system. DNA expression changes in these networks may have an important influence on Alzheimer's development.

This investigation highlights the importance of using non-linear methods to study DNA changes in Alzheimer's disease. Demonstrating that a deeper understanding of these genetic and molecular pathways could lead to earlier detection and innovative treatments for the disease.

1. Introduction

1.1 Alzheimer's Disease

Alzheimer's disease (AD) is a neurodegenerative disease characterised by cognitive decline and the accumulation of amyloid β-plaques and neurofibrillary tangles composed of hyperphosphorylated tau (1). AD accounts for 60-70% of dementia cases (2) and affects 57 million people worldwide, a prevalence which is expected to increase to 152 million by 2050 (3). In 2019, the economic burden of AD and related dementias was estimated at $1.3 trillion annually (4) and is expected to rise to $2 trillion annually by 2050 (5). The importance and economic benefits of AD research cannot be overstated; discovering novel biomarkers and developing treatment options can potentially ameliorate its global, debilitating effect.

1.2 Braak Staging

Tau is a microtubule-associated protein that stabilises neuronal cytoskeletal structures. In AD, tau becomes hyperphosphorylated and forms neurofibrillary tangles (NFTs), which disrupt axonal transport, synaptic function, and neuronal survival (6). The distribution of tau pathology forms the basis of the neuropathological framework Braak staging. Braak staging classifies AD progression according to the spread of NFTs from the entorhinal cortex to limbic regions and eventually neocortical areas. In early Braak stages (I-II) NFTs are restricted to the entorhinal cortex. In moderate Braak stages, tau pathology is present in the hippocampus and limbic areas (Stages III-IV). Finally, NFTs reach the pre-frontal cortex and association cortices (Stages V-VI) accompanied by substantial epigenetic dysregulation (7, 8).

1.3 DNA Methylation

DNA methylation is an epigenetic modification defined by the addition of a methyl group ($-CH_3$) to cytosine residues in CpG dinucleotides. CpG sites are characterised by a cytosine residue bonded to guanine via a phosphate group in a 5' to 3' direction; they are essential in epigenome-wide association studies (EWAS) studies because methylation can impair transcription factor binding, silencing the gene associated with the methylated CpG site (9). Methylation patterns have been widely implicated across multiple brain regions in AD (10, 11) as thousands of CpG sites related to AD pathology have been identified (12).

EWAS studies have identified widespread differentially methylated regions (DMRs) associated with Braak staging, suggesting that epigenetic mechanisms regulate tau pathology in AD progression (10, 11). While thousands of AD-related CpG sites have been identified the exact epigenetic mechanism of AD remains elusive (13), despite our understanding substantially advancing. Further identification of Braak-associated CpG sites and their specific impact on gene expression is essential to model AD progression entirely.

1.4 Limitations of Previous Literature

The study of epigenetics has yet to translate into effective therapies for AD despite having the potential to do so. This is due to the complex nature of DNA methylation, difficulties in proving causality, and distinguishing disease-driven epigenetic changes from age-related epigenetic drift (14). Furthermore, current treatment options, such as acetylcholinesterase inhibitors and NMDA receptor antagonists, provide only symptomatic relief and do not affect disease progression. All previous literature has used linear models, assuming a constant relationship between DNA

methylation and Braak staging. No previous literature has used non-linear modelling, and thus, expanding the search for non-linear associations in DNA methylation may uncover previously overlooked interactions, offering new insights into the epigenetic regulation of AD.

Non-linear modelling offers greater flexibility than linear models, aligning with the innate non-linearity of epigenetics and biological systems. Therefore, non-linear models have the potential to identify previously unassociated CpG sites (8). No study has modelled DNA methylation non-linearly in AD before, leaving a critical gap in our understanding. (4,5). Furthermore, non-linear feature selection methods can quantify the relative influence of methylation at specific CpG sites over Braak staging, providing deeper insight into which epigenetic changes are most relevant to AD progression.

Identifying highly correlated CpG sites could allow for earlier intervention through epigenetic screening, as epigenetic changes occur years before clinical symptoms appear. Blood-based DNA methylation signatures are being explored as non-invasive indicators of AD progression, potentially enabling earlier diagnosis and personalised disease monitoring (15). Since DNA methylation regulates tau, methylation modelling could provide novel disease-modifying pathways (16).

1.5 Aims

Therefore, this study aims to [1] identify non-linear associations between DNA methylation and AD pathology, [2] determine which methylation loci have the most influence on disease progression, and [3] assess the functional relevance of CpG through genetic enrichment analysis. We hypothesise that non-linear analysis will reveal novel insights into the epigenetic regulation of AD and identify previously overlooked CpG sites that contribute to disease progression.

2. Methods

2.1 Overview of ROS/MAP Dataset

The Religious Orders Study and Memory and Aging Project (ROS/MAP) is a longitudinal cohort study investigating neuropathological, genetic, and epigenetic factors underlying AD. The study comprises participants from two independent cohorts: the Religious Orders Study (ROS), which enrols Catholic nuns, priests, and brothers from across the United States, and the Memory and Aging Project (MAP), which recruits from retirement communities and senior housing facilities in the Chicago area. Both cohorts undergo annual cognitive and clinical assessments, with participants agreeing to post-mortem brain donation (17).

Post-mortem prefrontal cortex tissue was selected for DNA methylation analysis, as this region exhibits substantial epigenetic alteration in AD (10). DNA methylation levels were measured using the Illumina HumanMethylation450K BeadChip array, which quantifies methylation across more than 480,000 CpG sites genome-wide using site-specific probes.

The outcome variable in this study was Braak stage, a semi-quantitative measure of neurofibrillary tangle (NFT) burden used to assess AD severity. Braak staging categorises NFT pathology into seven stages (0-VI), with lower stages (I-II) indicating early pathology in the entorhinal cortex, intermediate stages (III-IV) reflecting spread to the limbic system, and advanced stages (V-VI) marking neocortical involvement and Braak stage 0 indicating no NFT presence.

2.2 Data Pre-processing and Normalisation

Before this study, ROS/MAP was regressed to minimise technical variability and ensure high data quality as per RG Smith et al. (10). Raw intensity files were loaded using the minfi and wateRmelon packages in R. Quality control measures were applied, including background signal intensity checks, detection p-value assessments, and bisulfite conversion efficiency evaluations. Samples with excessive signal deviation, detection p-values above 0.005, or bisulfite conversion rates below 80% were removed. Sex mismatch was assessed through principal component analysis (PCA), and discrepancies resulted in sample exclusion. Genetic concordance was verified using single nucleotide polymorphism probes, with samples exhibiting genetic correlation coefficients above 0.65 excluded to prevent confounding from duplicate or closely related samples.

Probes with detection p-values above 0.05 in more than 5% of samples were removed. Cross-hybridised probes and those containing single nucleotide polymorphisms at the CpG site were excluded to avoid artefacts from sequence variation. Following quality control, quantile normalisation was applied using wateRmelon's dasen function to standardise methylation β-values across samples. Surrogate Variable Analysis (SVA) was used to correct batch effects, and cell composition estimates were obtained using the Cell Epigenotype Specific Tool to adjust for neuronal and glial proportions in the prefrontal cortex samples. Residual methylation values were calculated by regressing out the effects of age, sex, and cell composition, ensuring that downstream analyses focused on methylation differences attributable to AD pathology rather than confounding variables.

2.3 Generalised Additive Models (GAMs)

To model the non-linear association between DNA methylation levels and Braak stage, Generalised Additive Models (GAMs) were implemented using the mgcv package in R. Unlike traditional linear models, GAMs allow for flexible curve fitting without assuming a predefined relationship (17). Braak stages were defined as the response variable, which was modelled using a Gaussian family and treated as a continuous variable. Methylation β-values were defined as the predictor variable. Thin-plate regression splines were applied to optimise smoothness without requiring explicit knot placement, and the smoothing parameter was estimated via Restricted Maximum Likelihood (REML) to balance flexibility and interpretability. The number of knots was set to k = 4 to allow moderate non-linearity without excessive model complexity. Model fit was assessed using the deviance explained and the fitted splines' effective degrees of freedom (EDF), which measure the proportion of variability in data the model explains and the flexibility of the model's fit, respectively.

Several machine learning models, such as feed-forward neural networks, deep neural networks, and support vector machines, were tested to determine the most suitable approach for modelling DNA methylation changes associated with AD pathology. Ultimately, GAMs were selected for their non-linearity, ability to visualise relationships, and interpretability.

2.4 Cross-Validation of GAMs

Ten-fold cross-validation was performed to evaluate model stability and generalisability. The smoothing spline complexity was controlled using k = 4. Cross-validation errors were recorded to assess predictive performance. The original parameters were retained as hyperparameter tuning did not significantly reduce cross-validation error.

2.5 Multiple Testing Correction

A quantile-quantile evaluation assessed whether the distribution of p-values followed expected patterns under the null hypothesis. Subsequently, the Benjamini-Hochberg (BH) method was used to control the False Discovery Rate (FDR). The adjusted p-value threshold was set at FDR < 0.1 to ensure enough probes were retained for downstream analyses. Probes meeting the FDR threshold were carried forward for further analysis. Genomic inflation factor (λGC) was calculated to assess potential inflation in test statistics. λGC values close to 1 indicate well-calibrated test statistics, whereas values much greater than 1 suggest potential inflation. Probes meeting the FDR threshold were carried forward for further analysis.

2.6 Epigenome Wide Association Study (EWAS) Catalogue

To determine whether CpG sites identified in our analysis had been previously associated with Alzheimer's disease (AD), we cross-referenced our significant probes with the EWAS Catalogue. This publicly available database compiles findings from published EWAS, allowing for comparisons of CpG sites across different studies. Each probe was queried against the catalogue to assess prior associations with AD or related neuropathologies (12).

2.7 Feature Selection

Feature selection was conducted using an Ensemble Feature Selection (EFS) approach to rank the most influential methylation loci associated with Braak stage. This method combined Random Forests, Elastic Net Regression ($\alpha = 0$, Ridge Regression), Gradient Boosting Machines (GBM), and Spearman's Rank Correlation to aggregate the strengths of each model. Random Forests and Gradient Boosting Machines were selected for their ability to model non-linear interactions, while Ridge Regression was used to address collinearity in methylation data by penalising correlated features proportionally meanwhile Spearman's Rank Correlation provided an independent ranking of features. Feature importance scores were normalised across models and aggregated into a final ranking.

2.8 Genetic Regional Enrichment Analysis Tool

After multiple testing corrections, significant probes were assigned to genes using the Genomic Regions Enrichment of Annotations Tool (GREAT) with the Human: GRCh37 (UCSC hg37) species assembly. This tool maps CpG sites to genetic regions using the basal plus extension method with the default parameters of proximal regions up to 5 kb upstream, 1 kb downstream, distal regions up to 1000 kb, and curated regulatory domains (19).

2.9 Functional Enrichment Analysis

Two independent enrichment analyses were conducted to investigate the biological relevance of significant DNA methylation changes. Following gene assignment, functional enrichment analysis will be performed using the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) to identify Gene Ontology (GO) terms and pathways from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) associated with significant loci (20, 21).

2.10 Fast Gene Set Enrichment Analysis (fGSEA)

fGSEA was conducted using the GAM-significant gene list annotated by GREAT (version 4.0.4) and pre-ranked by ensemble feature importance scores. The fgseaMultilevel function in RStudio

aggregated gene sets from the Molecular Signatures Database, including Reactome pathways and Gene Ontology (GO) Biological Processes. Pathways with fewer than the minimum gene size (>20 for Reactome Pathways, >90 for GO Biological Processes) were excluded to focus on larger, more biologically meaningful pathways. Statistical significance was determined using False Discovery Rate (FDR) correction (FDR < 0.05).

The Normalised Enrichment Score (NES) was used to standardise enrichment scores across different gene sets, accounting for variations in gene set size. NES is calculated by normalising the raw enrichment score (ES) against the mean ES of randomly permuted gene sets. A positive NES indicates pathway upregulation, while a negative NES suggests downregulation in the ranked gene list.

2.11 Protein-Protein Interaction Network

Protein-protein interaction (PPI) networks were constructed using the STRING database (v12.0) to analyse functional connectivity between identified proteins. The network was generated using a confidence-based interaction approach, where line thickness represents the strength of data support. The minimum required interaction score was set to 0.85 (high confidence) to ensure robust associations. Active interaction sources included text mining, experiments, databases, co-expression, neighbourhood, gene fusion, and co-occurrence. Stochastic flow was applied in the Markov Cluster Algorithm (MCL) to simulate random walks through the network, allowing for the identification of tightly connected protein clusters based on probabilistic flow dynamics (22).

The methodological overview is shown in Figure 1 to visualise the analysis pipeline.

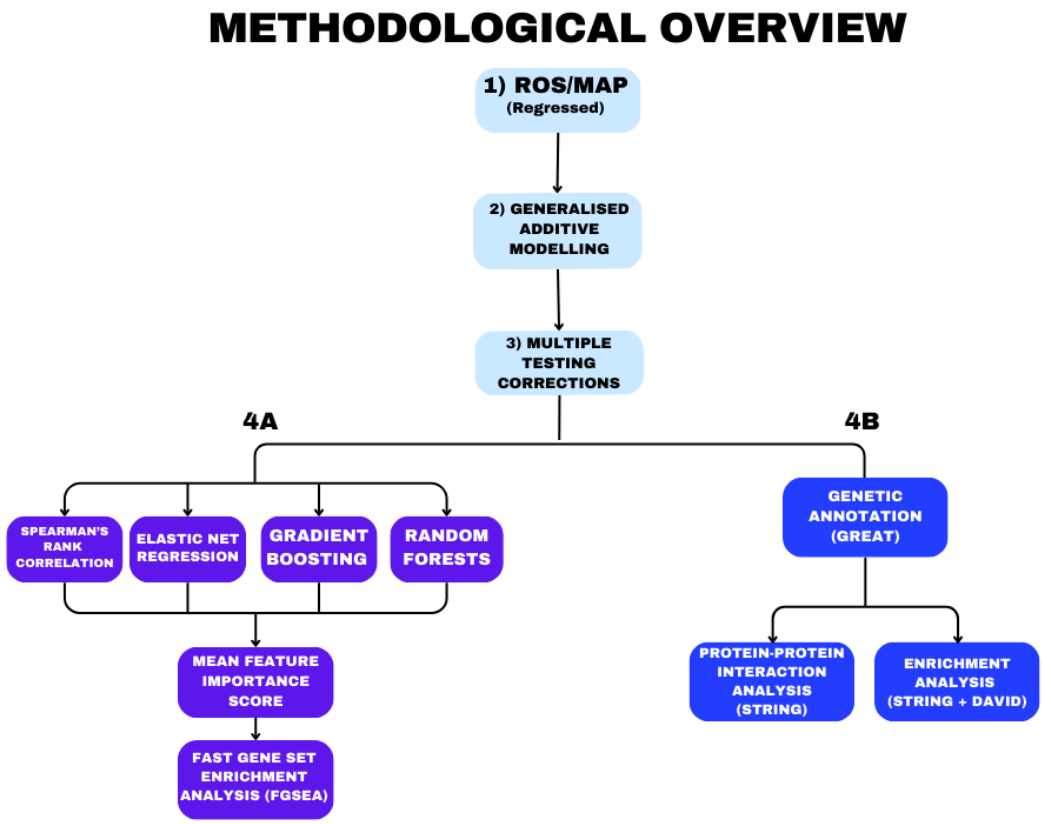

Figure 1: Methodological overview of the analytical pipeline. This diagram summarises the analytical workflow for investigating non-linear methylation changes associated with AD pathology. 1) Regressed ROS/MAP dataset. 2) Generalised Additive Models (GAMs) are applied to model methylation values against Braak stage. 3) Multiple testing corrections are used to correct the FDR. 4) Split into two parallel analytical branches: Panel 4A: Ensemble Feature Selection (EFS) combines four independent methods, Spearman's rank correlation, Elastic Net Regression, Gradient Boosting, and Random Forests to rank CpG sites by importance. Scores are averaged to calculate a Mean Feature Importance Score used in Fast Gene Set Enrichment Analysis (fGSEA). Panel 4B: Significant CpG sites are annotated to genes using the Genomic Regions Enrichment of Annotations Tool (GREAT). These genes are then analysed for functional relationships using Protein-Protein Interaction (PPI) analysis via STRING and pathway enrichment via DAVID.

3. Results

After quality control, pre-processing, and the removal of NA values, 711 samples and 412,999 CpG sites remained available for analysis. GAMs for every probe were modelled and subsequently cross-validated. The average cross-validation error across probes was 1.608, which was not significantly reduced after hyper-tuning (1.608 vs. 1.606). In total, 42,578 significant probes were identified by GAMs at the $p<0.05$ significance level.

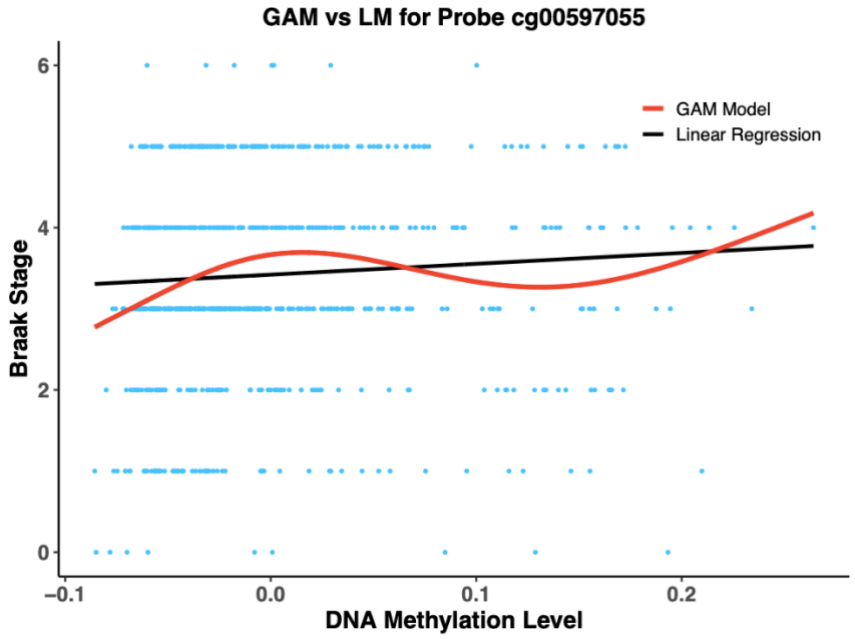

Figure 2: Linear model fails to capture significant association between DNA methylation and AD pathology at probe cg00597055. This plot compares Generalised Additive Model (GAM) and linear regression fits for the relationship between DNA methylation and Braak stage at probe cg00597055 of the Illumina HumanMethylation450K BeadChip array. Each blue point represents an individual sample, with DNA methylation level on the x-axis and Braak stage on the y-axis. The red line shows the GAM fit, while the black line shows the linear regression model. The linear model failed to detect a significant association ($p = 6.75e{-2}$), while the GAM revealed a significant non-linear relationship ($p = 6.23e{-5}$, $F = 8.08$, $EDF = 2.89$), demonstrating the added sensitivity of non-linear modelling for detecting epigenetic associations in Alzheimer's disease.

Figure 2 demonstrates that GAMs captured significant associations between DNA methylation and AD pathology at specific CpG sites, which traditional linear regression did not. Demonstrating the utility of non-linear models and the gap in AD literature which this investigation addresses.

Table 1: GAM Statistics for the Top-Ranked Probes by P-Value			
Probe	**Raw P-Value**	**F-Value**	**EDF**
cg01373852*	1.17E-14	24.56	2.98
cg05731218	1.19E-09	18.21	2.97
cg05684907*	1.85E-09	15.9	2.49
cg25307371*	2.16E-09	16.97	2.81
cg15016740*	3.18E-09	15.2	2.97
cg14315232*	3.87E-09	15.66	2.9
cg12307200	8.16E-09	19.88	2.09
cg04157161	1.05E-08	34.95	1
cg24862510*	2.19E-08	14.64	2.91
cg00979931	4.89E-08	30.92	1
cg01124843*	6.05E-08	13.69	2.95
cg13390284	9.40E-08	14.15	2.48
cg15033653	1.33E-07	28.93	1
cg22962123	1.44E-07	28.94	1

Note: EDF = Estimated Degrees of Freedom, quantifying spline flexibility. EDF > 1 indicates a non-linear association; EDF = 1 suggests linearity. *Probes marked with an asterisk have not been previously associated with Alzheimer's disease (EWAS Catalog). All p-values, F-values, and EDF values are rounded to two decimal places.

The variation in Estimated Degrees of Freedom (EDF) highlights GAMs' flexibility in capturing non-linear relationships. For example, cg01373852 (EDF = 2.98) and cg15016740 (EDF = 2.97) demonstrated non-linearity, while some probes, such as cg15033653 (EDF = 1), retained linear effects. High F-values indicate that the smoothing spline fits the data well and captures meaningful variations in methylation patterns associated with the Braak stage. However, the average cross-validation error (1.608) suggests that GAMs are inaccurate predictive models.

3.1 Multiple Testing Corrections

Since we fitted GAMs to 412,999 probes, we must control the false discovery rate. We determined that Bonferroni corrections were overly conservative and did not retain enough probes for downstream analysis, so we used Benjamini-Hochberg. Figure 3 demonstrates the need for an FDR-based approach to balance false discovery control with signal retention.

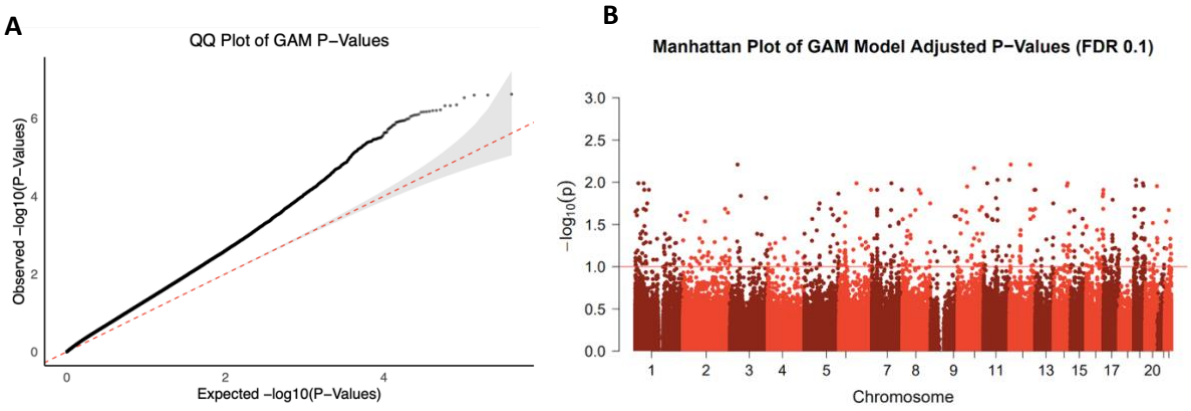

Figure 3: Distribution and significance of p-values from Generalised Additive Models (GAMs).
A) Quantile-Quantile (QQ) plot comparing observed and expected $-\log_{10}$(p-values) from GAMs across 412,999 probes. The

red dashed line represents the expected null distribution, and the grey shaded area shows the 95% confidence interval under the null hypothesis. Black points represent observed $-\log_{10}$(p-values). Deviation above the confidence interval indicates enrichment of low p-values, suggesting the presence of true associations. B) Manhattan plot of $-\log_{10}$(BH-adjusted p-values) for all CpG sites analysed using GAMs, plotted by chromosomal position. Each point represents a CpG site. The red horizontal line denotes the false discovery rate (FDR) threshold of 0.1. A total of 534 CpG sites exceeded this threshold, indicating statistically significant associations with Braak stage progression.

Panel B highlights the large subset of CpG sites exhibiting significant methylation changes associated with the Braak stage. After BH correction, 534 CpG sites crossed the p>0.1 significance threshold, 267 of which were not previously associated with AD (12).

3.2 Ensemble Feature Selection

Ensemble feature selection successfully distinguished the influence of each probe's effect on Braak staging. Table 2 presents the top 10 probes ranked by feature importance. GREAT successfully associated these CpG sites with genes.

Table 2: Top 10 Probes Ranked by Feature Importance

Probe	Average Importance Score	Associated Gene(s)
cg24862510*	0.47	PBX1; NUF2
cg12307200	0.45	TPRG1; LPP
cg02827029	0.44	RNF34; KDM2B
cg00059161*	0.44	SI; SLITRK3
cg00002033	0.43	LRFN1; IFNL1
cg03976468*	0.43	ADORA2A; SPECC1L
cg25403721*	0.41	TBX5; TBX3
cg00233028	0.38	TSPAN18; CD82
cg00574530*	0.37	NEUROG1
cg19011001	0.37	ITPK1; CHGA

Note: Probes marked with an asterisk have not been previously associated with Alzheimer's disease, according to the EWAS Catalog (ewascatalog.org). Average importance scores were derived from an ensemble feature selection method combining Random Forests, Elastic Net Regression, Gradient Boosting, and Spearman's Rank Correlation. Values are rounded to two significant figures.

Among the top 10 most influential probes identified by GAMs, 5 have no prior AD association. Furthermore, Gene-level annotation revealed their associated genes, including SI, TBX5, and NEUROG1, which were not previously associated with AD. The proportion of variance explained (R^2) of probe subsets ranked by feature importance was calculated to evaluate how well different subsets of probes explain variance in Braak staging. The top 10 and top 50 probe subsets explained relatively high variance per probe in Braak staging (R^2 = 0.12, 0.2, respectively) compared to the top 300 and top 534 probe subsets (R^2 = 0.31, 0.34, respectively), validating our ensemble feature selection approach. As the variance explained increased as more probes were included, all 534 probes were used in the genetic enrichment analysis.

3.3 Genetic Enrichment Analysis

fGSEA determined just one significant Reactome pathway (Developmental Biology, p = 0.0022, FDR = 0.0433, NES = 1.802, gene set size = 49) and two significant GO biological pathways: Positive Regulation of RNA Metabolic Process (p = 0.0053, FDR = 0.0372, NES = 1.6039, gene set size =

97) and Positive Regulation of Nucleobase-Containing Compound Metabolic Process (p = 0.0139, FDR = 0.0489, NES = 1.5159, gene set size = 102). All pathways were enriched for transcriptional and metabolic regulation genes, notably HOXA2 and HOXA3. Meanwhile, DAVID identified 39 significantly enriched biological processes, molecules, components, and pathways after FDR correction, while STRING identified 224. GREAT successfully mapped 534 regions to 710 genes, including previously associated genes such as ESR1, BIN1 and ANK1.

Table 3: DAVID Cluster analysis reveals highly enriched ontological group

Enrichment Score: 5.93

Gene Ontology Term	Count	Fold Enrichment	Adjusted P-Value
Chromatin	80	2.23	1.51E-08
Positive regulation of transcription by RNA polymerase II	75	1.89	4.91E-04
DNA-binding transcription factor activity, RNA polymerase II-specific	79	1.85	1.53E-04
Regulation of transcription by RNA polymerase II	94	1.71	8.69E-04
Sequence-specific double-stranded DNA binding	42	2.25	2.16E-03
RNA polymerase II cis-regulatory region sequence-specific DNA binding	70	1.76	5.31E-03
DNA-binding transcription activator activity, RNA polymerase II-specific	37	2.25	9.59E-03
DNA-binding transcription factor activity	42	2.11	0.01

Note: This table presents results from the DAVID cluster analysis, highlighting a highly enriched Gene Ontology (GO) group with an overall enrichment score of 5.93. The listed GO terms are associated with chromatin organisation and transcriptional regulation. Columns show the number of genes associated with each GO term (Count), the degree of overrepresentation relative to the background (Fold enrichment), and statistical significance after multiple testing corrections with the Bonferroni method (Adjusted p-value). All values are rounded to two decimal places.

Cluster analysis identified a highly enriched functional group with an overall enrichment score of 5.93, highlighting processes related to chromatin organisation, transcriptional regulation, and DNA-binding activity. The most significantly enriched term within this cluster was chromatin organisation (Fold Enrichment = 2.23, Adjusted P = 1.52E-08), indicating a strong association between epigenetic modifications and transcriptional regulation in AD. Several terms related to RNA polymerase II-mediated transcription were also highly enriched, including positive regulation of transcription by

RNA polymerase II (Fold Enrichment = 1.89, Adjusted P = 4.91E-04) and regulation of transcription by RNA polymerase II (Fold Enrichment = 1.71, Adjusted P = 8.69E-04).

Additionally, multiple terms associated with DNA-binding transcription factor activity were significantly enriched, including RNA polymerase II-specific transcription factor activity (Fold Enrichment = 1.85, Adjusted P = 1.53E-04) and sequence-specific double-stranded DNA binding (Fold Enrichment = 2.25, Adjusted P = 2.16).

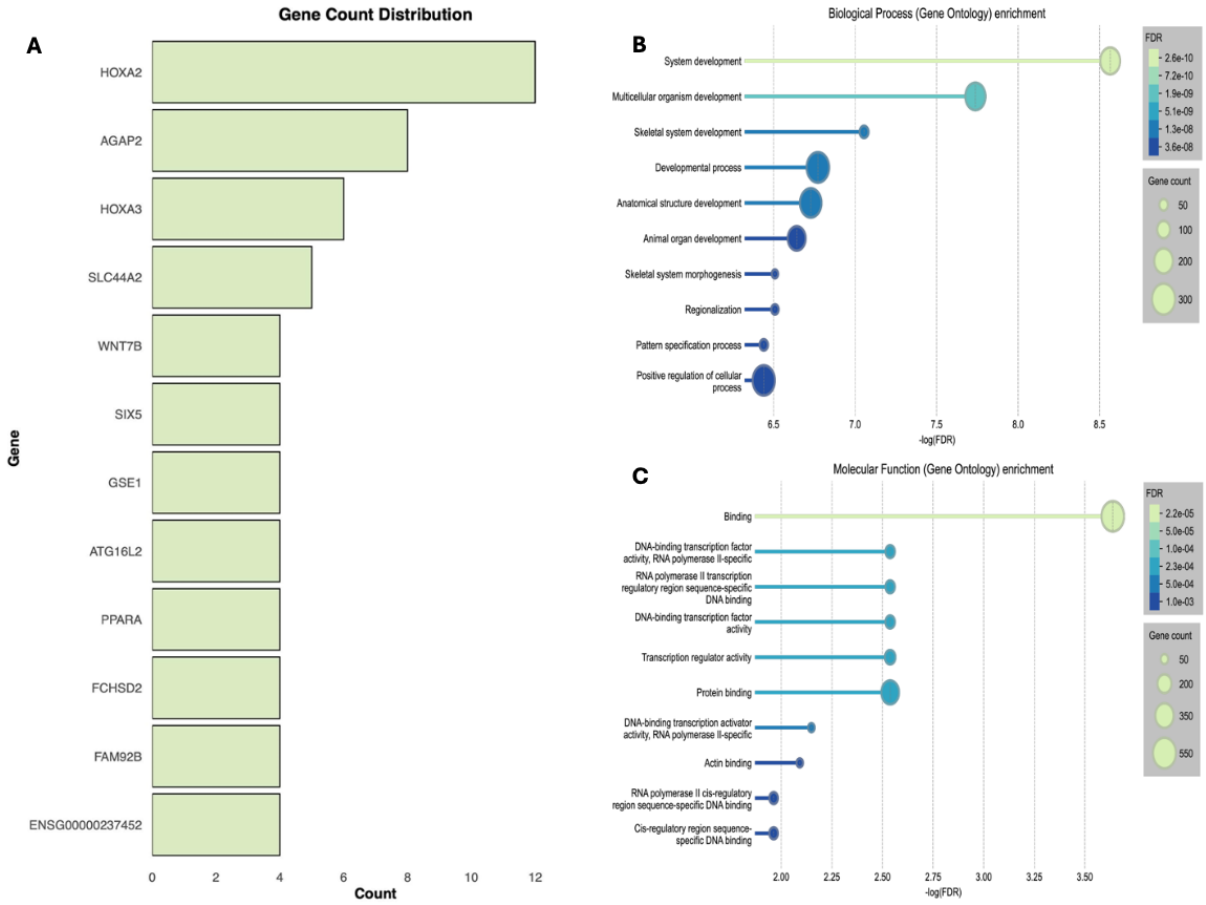

Figure 4. Visualisation of enriched genes and Gene Ontology (GO) terms associated with significant probes identified via Generalised Additive Models (GAMs). A) Gene count distribution showing the most frequently annotated genes (n ≥ 4) based on GREAT annotations of significant CpG sites. HOXA2, AGAP2, and HOXA3 were the most frequently annotated. B) Biological process enrichment analysis (Gene Ontology terms), highlighting enrichment in system development, rhombomere development, axonogenesis, and transcriptional regulation. C) Molecular function enrichment analysis, showing overrepresentation of DNA-binding transcription factor activity, RNA polymerase II-specific activity, and chromatin binding. All GO enrichments were conducted using STRING functional annotation tools and are shown with false discovery rate (FDR)-adjusted p-values < 0.05. Circle size indicates gene count, and colour intensity reflects significance level.

Gene Distribution Analysis (Figure 4A)

The most frequently enriched genes post-annotation were HOXA2 (12 occurrences), AGAP2 (8 occurrences), HOXA3 (7 occurrences), and SLC44A2 (5 occurrences).

Biological Process Enrichment (Figure 4B)
Several biological processes were significantly enriched, with system development showing the highest significance (Gene Count = 222, Strength = 0.21, signal = 0.74, FDR = 2.72E-09). Rhombomere development reported the highest strength (Gene Count = 5, Strength = 1.3, Signal = 0.58, FDR = 0.0057), four of which were homeobox genes, and embryonic skeletal system development reported the highest signal (Gene count = 24, Strength = 0.74, Signal 1.12, FDR = 8.17e-07) 13 of which were homeobox genes.

Molecular Function Enrichment (Figure 4C)
The molecular function analysis identified binding activity as the most significantly enriched function (Count = 527, Strength, 0.06, Signal = 0.43, FDR = 2.6E-10). DNA-binding transcription repressor activity reported the highest strength (Count = 28, Strength = 0.37, Signal = 0.36, FDR = 0.0182) and RNA polymerase II specific DNA-binding transcription activity reported the highest signal (Gene count = 85, Strength = 0.24, Signal = 0.43, FDR = 2.9E-2).

3.4 Pathway Modelling

DAVID revealed three significantly enriched KEGG pathways: the Ras signalling pathway (Count = 22, Enrichment = 2.67, FDR = 2.1E-2), the Rap1 signalling pathway (Count = 19, Enrichment = 2.59, FDR = 4.2E-2) and the signalling pathways regulating the pluripotency of stem cells (Count = 15, Enrichment = 3.01, FDR = 4.2E-2). The latter has not yet been associated with AD; its pathway is shown below in Figure 5.

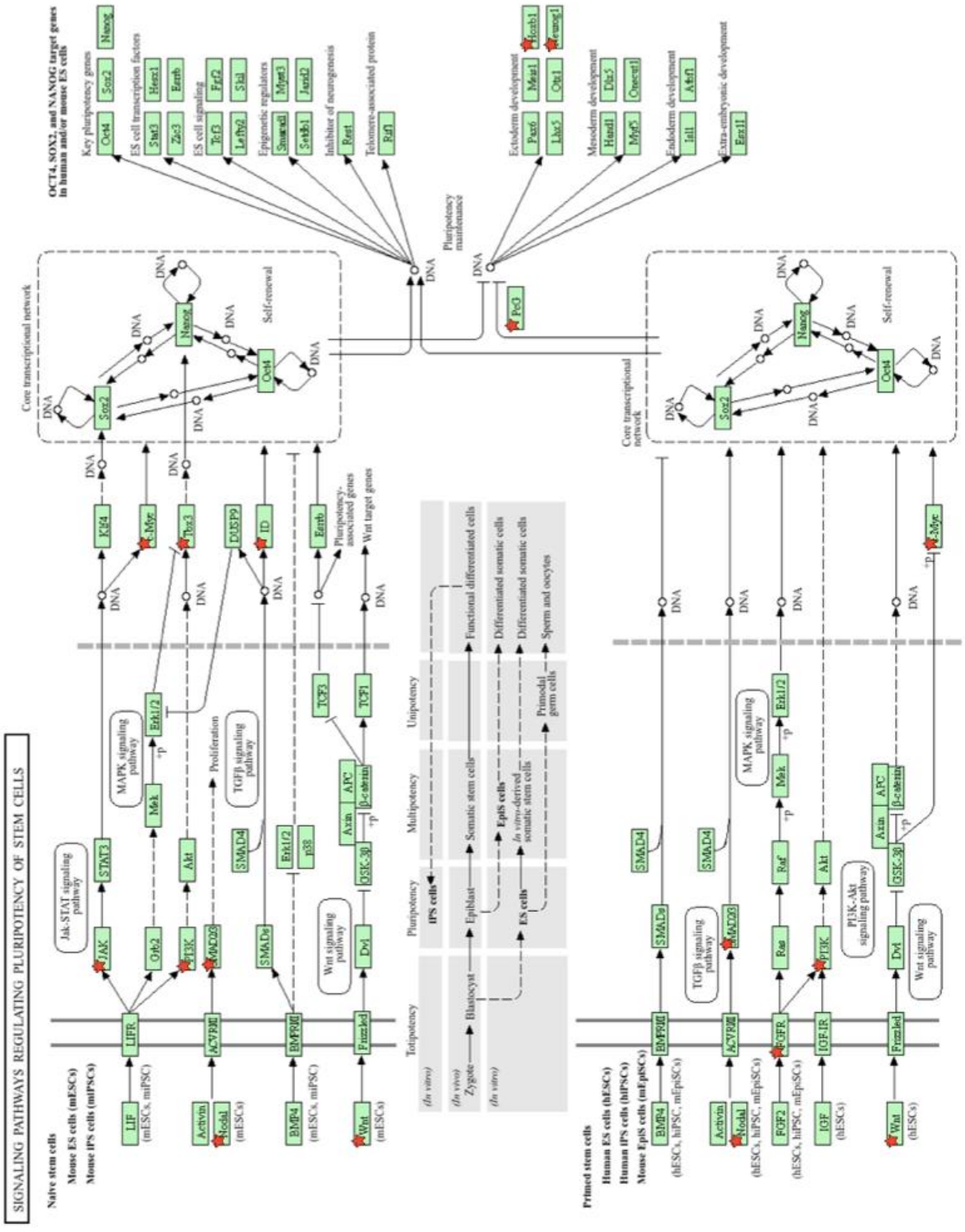

Figure 5. KEGG signalling pathways regulating pluripotency of stem cells, significantly enriched in the gene list derived from GREAT annotations of GAM-significant probes. This pathway was identified using DAVID functional annotation and visualised using KEGG. Genes marked in green were present in the input list. Enrichment statistics: Fold Enrichment = 3.01, $p = 4.37 \times 10^{-4}$, FDR = 0.0419. The pathway includes signalling cascades involved in stem cell maintenance and differentiation, such as JAK/STAT, TGF-β, and PI3K-Akt pathways.

The signalling pathways regulating the pluripotency of stem cells were significantly enriched (Count = 15, Fold Enrichment = 3.01, FDR =4.19E-2). The genes responsible for this enrichment included stem cell maintenance and differentiation regulators, notably JAK1, SMAD2, and PIK3CD. Furthermore, genes not previously associated with AD, such as NEUROG1, HOXA1, and HOXB1, are implicated in this pathway; more specifically, HOXB1 and NEUROG1 are implicated in ectoderm development (Figure 6).

3.5 Protein-Protein Network Analysis

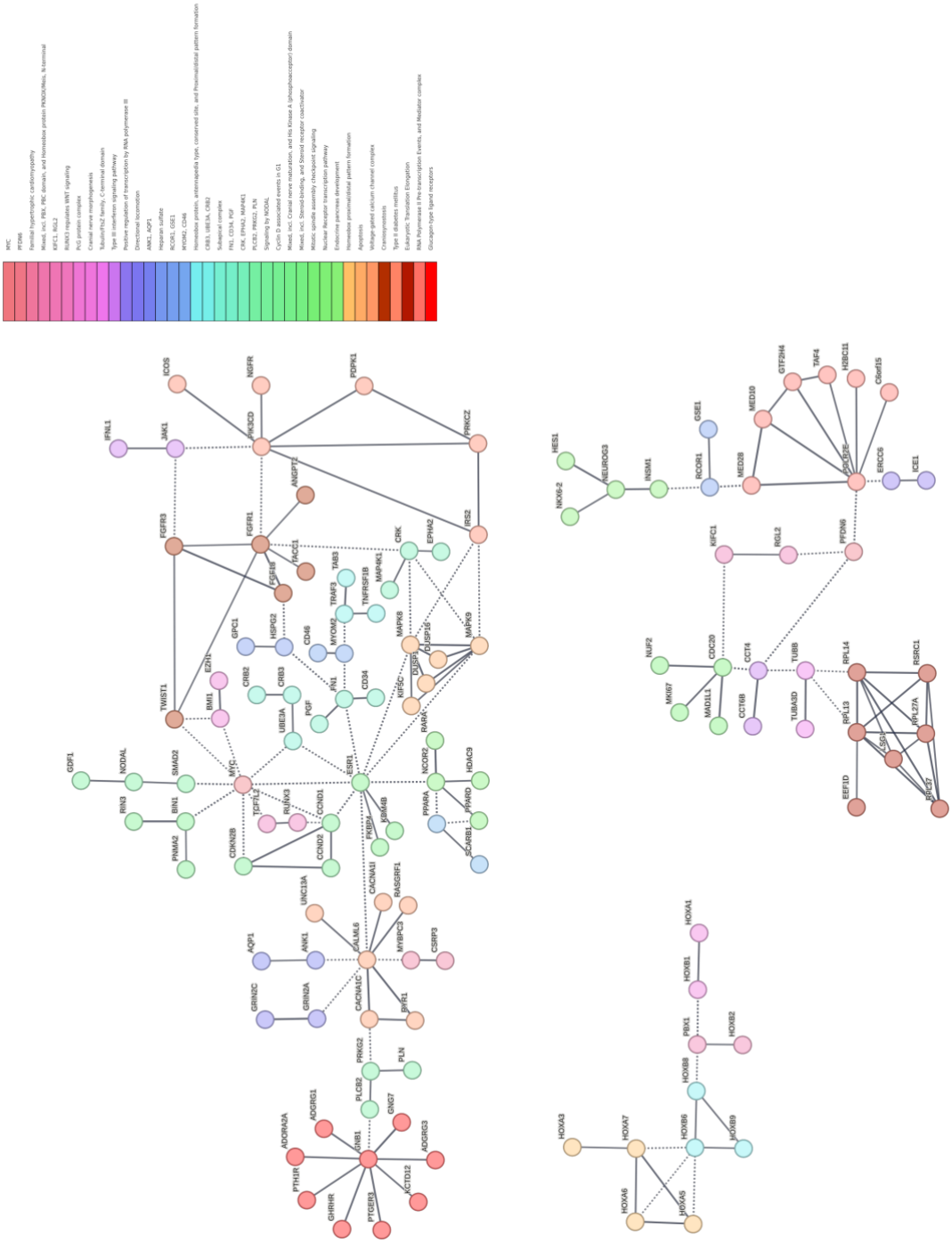

Figure 6. High-confidence protein-protein interaction (PPI) network constructed from the GREAT-annotated gene list using STRING. The network includes 705 nodes and 224 edges, with an average node degree of 0.64 and an average local clustering coefficient of 0.22. Only interactions with a STRING confidence score ≥ 0.85 (high confidence) are shown. Protein clusters were identified using the Markov Clustering (MCL) algorithm with a stochastic flow-based approach (inflation factor = 3). Only clusters containing more than seven proteins are displayed. Solid edges represent experimentally validated physical interactions, while dashed edges indicate predicted functional associations based on co-expression, co-occurrence, or curated databases.

The PPI enrichment (p=3.91E-3) indicates that the network exhibits statistically significant functional connectivity, exceeding the expected 186 edges for a randomly selected set of proteins. This suggests that the proteins involved share biologically meaningful interactions relevant to neurodegenerative processes in AD.

Three clusters with more than seven proteins were identified. Notably, within the HOX cluster (bottom left), HOXB1 is present, a protein that also appears in the signalling of pathways regulating the pluripotency of stem cells (Figure 5). Furthermore, HOXA7 and HOXB2 are present and have not previously been implicated in AD. Within the HOX cluster, HOXB6 exhibited moderate connectivity (node degree = 5), whereas other cluster members formed more discrete sub-networks.

A highly interconnected central cluster of 82 proteins previously implicated in AD was identified. Several genes within this cluster exhibited high connectivity, suggesting their importance in network stability and regulation. ESR1 had the highest connectivity (node degree = 10), followed by GNB1, MYC, and CALML6, (node degree = 9. Other highly connected proteins included MAPK8 and MAPK9 (node degree = 7), both members of the mitogen-activated protein kinase (MAPK) signalling cascade. BIN1 and ANK1 were also present, as was ADORA2A, a gene annotated to cg03976468, which had the sixth-highest mean feature importance (0.43).

4. Discussion

4.1 Enrichment of Homeobox Genes

GAMs identified many associations between DNA methylation and Braak stage in probes annotated to a subset of homeobox genes called HOX genes. HOX genes are crucial in neurodevelopment and epigenetic regulation (23). HOX genes are associated with neural differentiation, segmental identity, and synaptic specificity (24). During early brain development, they dictate hindbrain segmentation and neuronal fate (25, 26), but it has been suggested that they become dysregulated in AD (16). Our findings repeatedly identified HOX genes within enriched pathways, annotated to CpG sites with high feature importance, and in novel signalling pathways representing their broad implication in modulating epigenetic landscapes relevant to AD.

Beyond their developmental roles, molecular function and cluster analyses reinforce the transcriptional regulatory influence of these genes. Enrichment of binding activity, particularly DNA-binding transcription repressor activity, suggests a role in gene silencing and chromatin remodelling. This aligns with the identification of chromatin organisation as a highly enriched term, indicating a strong association between epigenetic modifications and transcriptional regulation in AD.

Additionally, multiple processes related to RNA polymerase II-mediated transcription were significantly enriched, further supporting the role of transcriptional regulation in neuronal function and disease pathology, and reinforcing the hypothesis that epigenetic dysregulation of transcriptional control is a key contributor to disease progression.

Process enrichment further supports the role of HOX gene-associated pathways in early neurodevelopment and later neurodegenerative processes. CpG sites linked to these genes may serve as epigenetic switches, influencing neuronal resilience and disease progression in aged and diseased states. Epigenetic modifications within these genes may exert long-term effects on brain function and AD susceptibility, reinforcing the theory that early epigenetic influences could contribute to neurodegenerative processes later in life. However, whether these methylation patterns are established prenatally or emerge in response to neurodegenerative pathology remains uncertain.

Since we adjusted for age and cell-type proportions, HOX gene methylation changes are likely linked to AD pathology rather than just ageing. However, we cannot entirely exclude all confounding factors (e.g., medication effects, epigenetic drift, comorbidities) (27).

4.2 PBX1 and NUF2

cg24862510 was identified as the CpG site with the highest feature importance, suggesting a strong association with Braak stage. This site, annotated to PBX1 and NUF2, has not previously been implicated in AD. PBX1 encodes a homeobox transcription factor essential for neuronal development, differentiation, and regional patterning during neurogenesis, suggesting it may influence neuronal resilience or vulnerability in neurodegeneration (28).

The high feature importance of cg24862510 and the location of PBX1 as a central node in the HOX gene cluster (Figure 6) suggest that cg24862510's methylation substantially influences the expression of homeobox genes. NUF2, however, was identified in the second-largest cluster of the PPI network, implying that methylation at cg24862510 influences Braak stage via two distinct mechanisms.

Given that this CpG site is located within a CpG island and deoxyribonuclease I hypersensitive site, methylation at this locus could alter the ability of PBX1 to regulate homeobox genes involved in synaptic maintenance and neuronal plasticity. Given the established role of homeobox genes in neurodevelopment, PBX1 expression might contribute to neurodegeneration by dysregulating neuronal differentiation pathways. Future research should investigate the functional consequences of cg24862510 methylation on the expression of PBX1 and NUF2 and their downstream targets in AD.

4.3 Signalling Pathways Regulating Pluripotency of Stem Cells

Identifying HOXB1 and its role in the signalling pathways regulating the pluripotency of stem cells, specifically ectoderm development, is highly relevant. Dysregulation of stem cell maintenance pathways may impair the brain's regenerative capacity, potentially contributing to neuronal loss and synaptic dysfunction in AD. If these pathways become silenced due to hypermethylation, the brain's ability to repair damaged networks or generate new neurons may be compromised, accelerating disease progression. This novel enrichment suggests that AD-associated methylation changes

interfere with transcriptional systems controlling adult neurogenesis, a process increasingly associated with neurodegenerative disease (29).

The identification of NEUROG1 in this pathway underscores its significance in neurodevelopment and neuronal lineage specification. Its associated CpG site (cg00574530) had the ninth-highest mean feature importance, suggesting that methylation at this site substantially influences Braak stage. NEUROG1 encodes a transcription factor essential for neuronal differentiation, particularly transitioning from neural progenitor cells to post-mitotic neurons (30). Therefore, changes in cg00574530 methylation could disrupt NEUROG1 expression, impairing the differentiation of progenitor cells into neurons and reducing the brain's ability to replace lost neurons.

Future research should investigate the functional consequences of methylation on NEUROG1 expression, particularly its impact on neuronal differentiation and survival. Since NEUROG1 is essential for neural fate commitment, its epigenetic dysregulation may contribute to neuronal regeneration deficits, further exacerbating AD neurodegeneration. Understanding how methylation at this site alters the role of NEUROG1 in neural differentiation may provide insights into potential epigenetic targets for therapeutic strategies to restore neurogenic capacity in neurodegenerative diseases.

IFNL1 was annotated to cg00002033, which has a high feature importance and has previously been linked to AD (12). Linked to JAK1 in the largest PPI cluster, IFNL1 encodes interferon lambda-1, a cytokine involved in the Janus Kinase-Signal Transducer and Activator of Transcription (JAK-STAT) signalling pathway, which is implicated in the signalling pathways regulating stem cell pluripotency (Figure 5).

Hypermethylation at cg00002033 affects IFNL1 expression, which can lead to altered immune responses in the brain. Altered IFNL1-JAK1 interactions may contribute to chronic neuroinflammation, a known driver of synaptic dysfunction and neuronal loss in AD (31). This suggests a link between neuroinflammation, stem cell maintenance, and neurodegeneration in AD.

Future research should focus on the effects of cg00002033 methylation on IFNL1 expression and its downstream signalling, particularly in the context of immune-mediated neurodegeneration and stem cell regulation. Understanding these epigenetic modifications may uncover novel therapeutic targets to modulate neuroimmune interactions in AD.

4.4 ADORA2A

Adenosine A2A receptor (ADORA2A) was associated with the sixth most influential CpG site by feature importance and showed significant functional enrichment across 76 GO terms. As a G-protein coupled receptor (GPCR), ADORA2A modulates synaptic plasticity, neuroinflammation, and energy metabolism, all implicated in AD (32). The presence of ADORA2A within the most extensive protein-protein interaction (PPI) cluster suggests that it acts as an essential regulatory node in the broader epigenetic landscape of AD and could have broad downstream transcriptional effects. The presence of BIN1 and ANK1 within this network, known risk factors of AD (10), adds credibility to our methods.

The association of cg03976468 with ADORA2A suggests a potential epigenetic regulatory role in AD. This CpG site is located within a SPECC1L-ADORA2A readthrough transcript, a nonsense-mediated

decay candidate. This indicates that it may not produce a stable protein. The site is situated within a CpG island, and the presence of H3K27Ac marks and DNase hypersensitivity signals suggest that this locus is located within an open chromatin region and may be accessible for transcription factor binding.

Additionally, predicted transcription factor binding sites (JASPAR, STAT2) further support the possibility that this region plays a role in transcriptional regulation. While ADORA2A is highly expressed in the brain (GTEx, RPKM = 8.26), and has been nominated as a potential therapeutic target (33, 34), the extent to which methylation at cg03976468 influences its expression remains unclear without functional validation. Future studies employing functional assays such as chromatin accessibility profiling, reporter assays, and CRISPR-based epigenetic editing will be necessary to determine whether methylation at this site modulates ADORA2A transcription and its role in AD pathology. Due to the lack of direct experimental evidence, this study can only propose a functional role for cg03976468 in ADORA2A regulation.

4.5 Methodological Limitations

While our study identified many significant non-linear methylation changes, several methodological limitations must be considered. They are listed below in approximate order of importance.

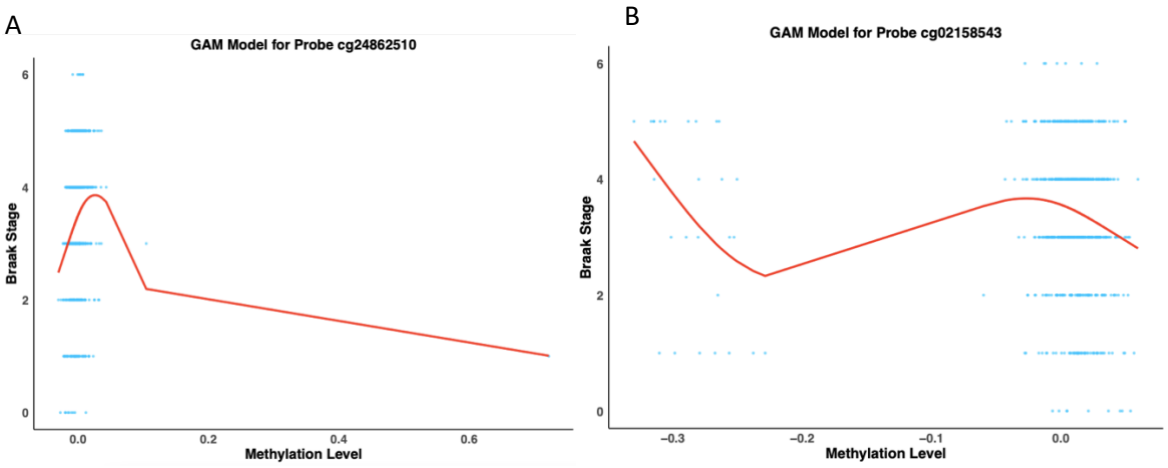

Figure 7. Unexpected modelling patterns in Generalised Additive Models (GAMs) for Alzheimer's disease methylation data. A) GAM fit for probe cg24862510, showing a sharp peak driven by dense data at low methylation levels and a flat fit across the remaining range. The bimodal distribution may reduce interpretability. B) GAM fit for probe cg02158543, illustrating a non-linear association with two visible methylation clusters.

1) Additional pre-filtering steps are required to remove artefacts; Figure 7 highlights the need for improved pre-filtering steps to enhance model stability and reduce artefactual signal. While surrogate variable analysis (SVA) was appropriately used to adjust for batch effects and unmeasured confounding (10), future studies could explore the use of Principal Component Analysis (PCA) as a complementary tool for tasks such as outlier detection, dimensionality reduction, or exploratory data cleaning.

2) Balancing the need to control p-value inflation with the risk of discarding relevant associations: Using a BH-adjusted <0.1 significance threshold increased our sensitivity to capture meaningful associations but also raised the risk of false positives. Our quantile-quantile plot (Figure 3A) suggests high p-value inflation. While additional attempts to reduce inflation through stricter correction methods may reduce this inflation, it risks discarding valid biological signals. Future studies should explore alternative pre-processing strategies, such as negative control-based methods, to control confounding variables more effectively (35).

3) Correlation, not Causation: Several highly important probes and genes are identified in this analysis (e.g., ADORA2A, PBX1, NUF2); however, this investigation cannot determine causality between methylation at the discussed CpG sites and AD pathology, our identified methylation changes could be a consequence of neurodegeneration, not a cause of it. Future studies should validate the CpG sites we have identified in independent datasets; leverage multi-omics approaches to elucidate the downstream effects of methylation and functional genomic editing to provide direct evidence of causality.

4) Sample and Study Design Limitations: The study is limited by its reliance on postmortem tissue, a single brain region, and one epigenetic mark (DNA methylation), thereby excluding potentially relevant spatial and epigenomic dimensions of AD. While surrogate variable analysis was used to control for potential post-mortem effects, the biological consequences of post-mortem interval may not be fully captured by statistical methods alone, potentially introducing noise into methylation values (36). Further investigations should compare non-linear methylation patterns across brain regions (37 and other epigenetic marks, such as histone modification or non-coding RNA-associated gene silencing (38, 39).

This study did not use a case-control framework, which is standard in epigenome-wide association studies (EWAS). Separating individuals into high (e.g., Braak V–VI) and low (e.g., Braak 0–II) pathology groups can highlight group differences and offer clearer biological contrasts (10). Although our continuous modelling approach is more sensitive to subtle methylation shifts across stages, a future hybrid framework incorporating continuous and categorical modelling could optimise sensitivity and specificity when identifying functionally relevant epigenetic changes.

As with every EWAS study, a larger sample size would allow for more statistical power when modelling. This would enable us to use more stringent significant levels and multiple testing correction methods, thereby increasing the confidence that significant probes have a biologically meaningful association with AD pathology (40).

5) Machine learning-based feature selection produced different scores across runs, leading to poor reproducibility, a known limitation of the technique (41). Future investigations with additional processing power can run the method multiple times and focus on consistently highly ranked probes across runs. Furthermore, their feature selection methods could use the full array of probes rather than just GAM-significant ones, as this investigation could have overlooked highly influential associations that were not GAM-significant.

6) GAMs mispredicted Braak staging by an average of 1.608 stages: The observed high CV error could be due to the semi-continuous nature of Braak staging. Previous literature has highlighted that continuous measures of tau pathology better reflect AD progression (42). Future investigations could

reduce and produce better predictive GAMs by using continuous measures of AD pathology, such as the global AD pathology score from the ROS/MAP dataset. Alternatively, the high cross-validation error observed could result from GAMs predicting Braak stages continuously, which leads to them predicting non-integer values which does not align with the nature of Braak staging. Future methods could continue to model GAMs non-linearly as we have done in this investigation but ensure ordinal regression techniques are used to predict Braak stage, potentially reducing CV error.

7) More complex methodological pipelines could model associations at a probe-by-probe level. For example, Figure 7B illustrates a bimodal methylation distribution with two clusters. In this case, k-means clustering would be a more appropriate way to visualise methylation at cg24862510 over a GAM. With additional computational power, more advanced methodological pipelines could iteratively fit different modelling techniques at a probe-by-probe level, using metrics such as CV error, Bayesian Information Criterion and Akaike Information Criterion to determine which model best explains the methylation pattern at each probe.

5. Conclusion

This study applied non-linear modelling to AD epigenetic data and identified 267 novel CpG sites associated with Braak staging, addressing a gap left by previous literature. Genetic enrichment and PPI analysis revealed that these CpG sites are functionally relevant, highlighting dysregulation in homeobox genes and stem cell pluripotency pathways. These findings support the hypothesis that early neurodevelopmental mechanisms are epigenetically altered in Alzheimer's and may contribute to later neurodegeneration.

This study demonstrates that non-linear models can uncover biologically meaningful methylation patterns that linear models overlook. This investigation advances our understanding of AD epigenetics and offers a framework for future epigenetic analysis. Thus, our hypotheses were supported: the non-linear GAM approach uncovered previously unrecognised methylation changes (Aim 1) and pointed to new biological pathways (Aim 3) that linear models had overlooked, whilst ensemble feature importance techniques determined the most influential CpG sites (Aim 2), fulfilling the study's objectives. Continued functional validation investigation will be essential to assessing causality and the biomarker potential of identified sites.

Code Availability

All scripts used to perform the data analysis can be found here
https://github.com/ArchieTruman/MethylationBraak

6. References

1. Montine TJ, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Dickson DW, et al. National Institute on Aging–Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. Acta Neuropathol. 2011 Nov;123(1):1–11. doi: 10.1007/s00401-011-0910-3.

2. McDiarmid AH, Gospodinova KO, Elliott RJR, Dawson JC, Graham RE, El-Daher MT, et al. Morphological profiling in human neural progenitor cells classifies hits in a pilot drug screen for Alzheimer's disease. Brain Commun. 2024;6(2). doi: 10.1093/braincomms/fcae101.

3. GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health. 2022 Jan;7(2):e105–25. doi: 10.1016/S2468-2667(21)00249-8.

4. Wimo A, Seeher K, Cataldi R, Cyhlarova E, Dielemann JL, Frisell O, et al. The worldwide costs of dementia in 2019. Alzheimers Dement. 2023 Jul;19(7):2865–73. doi: 10.1002/alz.12901.

5. Tay LX, Ong SC, Tay LJ, Ng T, Parumasivam T. Economic Burden of Alzheimer's Disease: A Systematic Review. Value Health Reg Issues. 2024 Mar;40:1-12. doi: 10.1016/j.vhri.2023.09.008. Epub 2023 Nov 14.

6. Medeiros R, Baglietto-Vargas D, LaFerla FM. The role of tau in Alzheimer's disease and related disorders. CNS Neurosci Ther. 2011 Oct;17(5):514–24. doi: 10.1111/j.1755-5949.2010.00177.x.

7. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82(4):239–59. doi: 10.1007/BF00308809.

8. Qazi TJ, Quan Z, Mir A, Qing H. Epigenetics in Alzheimer's disease: perspective of DNA methylation. Mol Neurobiol. 2018 Feb;55(2):1026–44. doi: 10.1007/s12035-016-0357-6.

9. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013 Jan;38(1):23–38. doi: 10.1038/npp.2012.112.

10. Smith RG, Pishva E, Shireby G, Smith AR, Roubroeks JAY, Hannon E, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. Nat Commun. 2021 Jun 10;12(1):3517. doi: 10.1038/s41467-021-23243-4.

11. Zhang L, Silva TC, Young JI, Gomez L, Schmidt MA, Hamilton-Nelson KL, et al. Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. Nat Commun. 2020 Nov 30;11(1):6114. doi: 10.1038/s41467-020-19791-w.

12. Battram T, Yousefi P, Crawford G, Prince C, Babei MS, Zheng J, et al. The EWAS Catalog: a database of epigenome-wide association studies [version 2; peer review: 2 approved]. Wellcome Open Res. 2022;7:41. doi: 10.12688/wellcomeopenres.17598.2.

13. Qin H, Liu J, Fang C, Deng Y, Zhang Y. DNA methylation: the epigenetic mechanism of Alzheimer's disease. IBrain. 2023 Aug 10;9(4):463–72. doi: 10.1002/ibra.12121.

14. Liu X, Jiao B, Shen L. The epigenetics of Alzheimer's disease: factors and therapeutic implications. Front Genet. 2018 Nov 30;9:579. doi: 10.3389/fgene.2018.00579.

15. Fransquet PD, Lacaze P, Saffery R, Phung J, Parker E, Shah R, et al. Blood DNA methylation signatures to detect dementia prior to overt clinical symptoms. Alzheimers Dement (Amst). 2020 Jul 9;12(1):e12056. doi: 10.1002/dad2.12056.

16. Smith RG, Hannon E, De Jager PL, Chibnik L, Lott SJ, Condliffe D, et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimers Dement. 2018 Dec;14(12):1580–8. doi: 10.1016/j.jalz.2018.01.017.

17. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis. 2018;64(S1):S161–89. doi: 10.3233/JAD-179939.

18. Hastie T, Tibshirani R. Generalized additive models. Stat Sci. 1986;1(3):297–310. Available from: http://www.jstor.org/stable/2245459.

19. Tanigawa Y, Dyer ES, Bejerano G. WhichTF is functionally important in your open chromatin data? PLoS Comput Biol. 2022 Aug 30;18(8):e1010378. doi: 10.1371/journal.pcbi.1010378.

20. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022 Jul 5;50(W1):W216–21. doi: 10.1093/nar/gkac194.

21. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. doi: 10.1038/nprot.2008.211.

22. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023 Jan 6;51(D1):D638–46. doi: 10.1093/nar/gkac1000.

23. Barber BA, Rastegar M. Epigenetic control of Hox genes during neurogenesis, development, and disease. Ann Anat. 2010 Sep 20;192(5):261–74. doi: 10.1016/j.aanat.2010.07.009.

24. Philippidou P, Dasen JS. Hox genes: choreographers in neural development, architects of circuit organization. Neuron. 2013 Oct 2;80(1):12–34. doi: 10.1016/j.neuron.2013.09.020.

25. Duverger O, Morasso MI. Role of homeobox genes in the patterning, specification, and differentiation of ectodermal appendages in mammals. J Cell Physiol. 2008 Aug;216(2):337–46. doi: 10.1002/jcp.21491.

26. Zheng C, Lee HMT, Pham K. Nervous system-wide analysis of Hox regulation of terminal neuronal fate specification in Caenorhabditis elegans. PLoS Genet. 2022 Feb;18(2):e1010092. doi: 10.1371/journal.pgen.1010092.

27. Lardenoije R, Iatrou A, Kenis G, Kompotis K, Steinbusch HW, Mastroeni D, et al. The epigenetics of aging and neurodegeneration. Prog Neurobiol. 2015 Aug;131:21-64. doi: 10.1016/j.pneurobio.2015.05.002.

28. Liu M, Xing Y, Tan J, Chen X, Xue Y, Qu L, et al. Comprehensive summary: the role of PBX1 in development and cancers. Front Cell Dev Biol. 2024;12:1442052. doi: 10.3389/fcell.2024.1442052.

29. Altuna M, Urdánoz-Casado A, Sánchez-Ruiz de Gordoa J, Zelaya MV, Labarga A, Lepesant JMJ, et al. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. Clin Epigenetics. 2019 Jun 19;11(1):91. doi: 10.1186/s13148-019-0672-7.

30. National Center for Biotechnology Information (NCBI). NEUROG1 neurogenin 1 [Homo sapiens (human)] – Gene ID: 4762 [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2025 Mar 23]. Available from: https://www.ncbi.nlm.nih.gov/gene/4762

31. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nat Rev Neurol. 2021;17(3):157–72. doi: 10.1038/s41582-020-00435-y.

32. National Center for Biotechnology Information (NCBI). ADORA2A adenosine A2a receptor [Homo sapiens (human)] – Gene ID: 135 [Internet]. Bethesda (MD): National Library of Medicine (US); [cited 2025 Mar 23]. Available from: https://www.ncbi.nlm.nih.gov/gene/135

33. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 2025 Jan 6;53(D1):D1243–9. doi: 10.1093/nar/gkae974.

34. Domenici MR, Ferrante A, Martire A, et al. Adenosine A2A receptor as potential therapeutic target in neuropsychiatric disorders. Pharmacol Res. 2019;147:104338. doi:10.1016/j.phrs.2019.104338.

35. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012 Jul;13(3):539–52. doi: 10.1093/biostatistics/kxr034.

36. Tsai PC, Bell JT. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. Int J Epidemiol. 2015 Aug;44(4):1429–41. doi: 10.1093/ije/dyv041.

37. Lunnon K, Smith R, Hannon E, De Jager PL, Srivastava G, Volta M, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. Nat Neurosci. 2014 Sep;17(9):1164–70. doi: 10.1038/nn.3782.

38. Faghihi MA, Modarresi F, Khalil AM, Wood DE, Sahagan BG, Morgan TE, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. Nat Med. 2008 Jul;14(7):723–30. doi: 10.1038/nm1784.

39. Klein HU, McCabe C, Gjoneska E, Sullivan SE, Kaskow BJ, Tang A, et al. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in aging and Alzheimer's human brains. Nat Neurosci. 2019 Jan;22(1):37–46. doi: 10.1038/s41593-018-0291-1.

40. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011 Jul;12(8):529–41. doi: 10.1038/nrg3000.

41. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36(11):1–13. doi: 10.18637/jss.v036.i11.

42. Moscoso, A., Wren, M.C., Lashley, T. et al. Imaging tau pathology in Alzheimer's disease with positron emission tomography: lessons learned from imaging-neuropathology validation studies. Mol Neurodegeneration 17, 39 (2022). doi: 10.1186/s13024-022-00543-x.