Sex-specific DNA methylation changes in Alzheimer's disease pathology

Lanyu Zhang[1], Juan I. Young[2,3], Lissette Gomez[3], Tiago C. Silva[1], Michael A. Schmidt[2,3], Jesse Cai[4], Xi Chen[1,5], Eden R. Martin[2,3], Lily Wang[1,2,3,5*]

[1] Division of Biostatistics, Department of Public Health Sciences, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

[2] Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

[3] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL 33136, USA

[4] Brentwood High School, 5304 Murray Ln, Brentwood, TN 37027

[5] Sylvester Comprehensive Cancer Center, University of Miami, Miller School of Medicine, Miami, FL 33136, USA

* To whom correspondence should be addressed.

Supplementary Text on Detailed Methods

1. Pre-processing of DNA methylation data
2. Single cohort analysis
3. Inflation assessment and correction
4. Meta-analysis
5. Identifying sex-specific changes
6. Enrichment and pathway analysis
7. Integrative methylation – gene expression analysis
8. Sex-specific mQTL analysis

1. Pre-processing of DNA methylation data

Quality control for CpG probes included removing probes with detection P-value < 0.01 in all samples of a cohort, those associated with cigarette smoking[1], or those having a single nucleotide polymorphism (SNP) with minor allele frequency (MAF) ≥ 0.01 present in the last 5 base pairs of the probe. Quality control for samples included restricting our analysis to samples with good bisulfite conversion efficiency (i.e., $\geq 88\%$) and principal component analysis (PCA). More specifically, PCA was performed using the 50,000 most variable CpGs for each cohort. Samples that were within $\pm$ 3 standard deviations from the mean of PC1 and PC2 were selected to be included in the final sample set. The quality controlled methylation datasets were next subjected to the QN.BMIQ normalization procedure[2] as previously described[3]. The recorded sex status of all samples matched those predicted based on methylation levels using getSex function in R package minfi. We removed batch effects by applying linear model methylation M value ~ methylation slide. For the ROSMAP cohort, we additionally included the variable "batch" that was available in the dataset to adjust for technical batches which occurred during data generation. The residuals (methylation residuals) estimated from this model were then used for subsequent analyses.

2. Single cohort analysis

To identify sex-specific DNA methylation changes in AD, we performed both a sex-stratified analysis and a sex-by-Braak stage interaction analysis for each of the four brain sample cohorts. In sex-stratified analysis, we tested methylation-Braak stage associations in female and male samples separately. In sex-by-Braak stage interaction analysis, we analyzed both female and male samples simultaneously and compared slopes for methylation-Braak stage associations in females and males.

In sex-stratified analysis, for each CpG, we applied the model methylation residuals ~ age at death + Braak stage + CETS[4] estimated neuron proportions to female samples and male samples separately. In sex-by-Braak stage interaction analysis, for each CpG, we applied the model methylation residuals ~ age at death + sex + Braak stage + sex*Braak stage + sex*age at death + CETS estimated neuron proportions to samples of both sexes.

For the analysis of differentially methylated regions (DMRs), we used the coMethDMR R package[5] to analyze 40,010 pre-defined genomic regions on the Illumina 450k arrays and identify co-methylated DMRs associated with Braak stage. The pre-defined genomic regions are regions on the Illumina array covered with clusters of contiguous CpGs where the maximum separation between any two consecutive probes is 200 base pairs. First, coMethDMR selects co-methylated sub-regions within these pre-defined contiguous genomic regions. Next, we summarized methylation M values within these co-methylated sub-regions using medians and tested them against AD Braak stage. The same linear models described for the analysis of CpGs were then applied to median value of each DMR.

3. Inflation assessment and correction

To assess inflation of the test statistics, we used quantile-quantile (QQ) plots of observed and expected distributions of P-values for each cohort. Because the conventional genomic inflation factor (lambda or λ

used interchangeably below) is dependent on the expected number of true associations, and in a typical EWAS it is expected that small effects from many CpGs might be associated with the phenotype, Iterson et al. (2017)[6] showed that the conventional genomic inflation factor would overestimate actual test-statistic inflation in EWAS. To estimate genomic inflations more accurately in EWAS, Iterson et al. (2017)[6] developed a Bayesian method that estimates and corrects inflation in EWAS based on empirical null distributions, which is implemented in the Bioconductor package bacon. We estimated genomic inflation factors using both the conventional approach and the *bacon* method. In addition, we also applied the *bacon* method to single cohort analysis results to obtain inflation-corrected effect sizes, standard errors, and p-values for each cohort.

4. Meta-analysis

The results of the *bacon*-corrected cohort-specific analysis were then combined using inverse-variance weighted meta-analysis models. The evidence for heterogeneity of study effects was tested using Cochran's Q statistic[7]. More specifically, the inverse-variance weighted fixed effects model was first applied to synthesize statistical significance from individual cohorts. Even though the fixed effects model for meta-analysis does not require the assumption of homogeneity[8], for the regions with nominal evidence for heterogeneity (nominal $P_{heterogeneity} < 0.05$), we also applied random effects meta-analysis[9] and assigned final meta-analysis P-value based on the random effects model. For each CpG (and for each DMR), we used the R package meta to obtain meta-analysis p-values for sex-by-Braak stage interaction, as well as Braak stage effect in female samples and male samples separately in sex-stratified analysis.

5. Identifying sex-specific changes

In sex-stratified analysis, we selected significant CpGs (or DMRs) with FDR < 0.05 in female samples or male samples separately. In sex-by-Braak stage interaction analysis, because the standard error of interaction effect `sex × Braak stage` is typically much larger than those for main `Braak stage` effects, the conventional approach for controlling false discovery rate often results in low power for discovering interaction effects[10]. Therefore, we used a stagewise analysis approach, previously proposed by van de Berge et al. (2017)[10], to help improve power in high-throughput experiments where multiple hypotheses are tested for each gene. More specifically, in the *screening step*, for each CpG (or DMR), we tested the global null hypothesis that there is methylation-Braak stage association in either male or female samples. Next, in the *confirmation step*, we considered three individual null hypotheses for each CpG (or DMR): (a) there is no methylation-Braak stage association in male samples; (b) there is no methylation-Braak stage association in female samples; and (c) the methylation-Braak stage associations in male samples and female samples are the same. For the CpGs (or DMRs) selected in the screening step, these three individual hypotheses were then tested while controlling family-wise error rate (FWER) as described in van de Berge et al. (2017)[10].

The stagewise analysis described above was implemented using the stageR package to identify CpGs (or DMRs) with significant differential methylation - Braak stage associations in females and males. In the screening step, we considered meta-analysis p-values for Braak stage in female samples and male samples (`p.meta.female`, `p.meta.male`), and used the minimum of these two meta-analysis p-values to represent

each CpG (or DMR). In the confirmation step, the parameter pConfirmation was defined using three p-values for each CpG (or DMR): p.meta.female, p.meta.male, and p.meta.interaction (meta-analysis p-value for sex × Braak stage).

6. Enrichment and pathway analysis

The probes on the Illumina 450k array are annotated according to their locations with respect to genes (TSS1500, TSS200, 5'UTR, 1stExon, gene body, 3'UTR, intergenic) or to CpG islands (island, shore, shelf, open sea). To understand the genomic context of sex-specific DNA methylation changes in AD, we compared the FDR significant methylation changes from sex-stratified analysis with different types of genomic features. As pathological AD-associated methylation changes can occur at both significant individual CpGs and significant DMRs, we considered the CpGs located at significant individual CpGs or within significant DMRs jointly in this analysis, by testing their over- and under-representation in different types of genomic features using Fisher's exact test. More specifically, the proportion of significant CpGs mapped to a particular type of genomic feature (e.g., CpG islands) (foreground) was compared to the proportion of CpGs on the array that mapped to the same type of genomic feature (background).

In addition, we used Fisher's test to assess enrichment of significant CpGs and DMRs in different chromatin states by comparing with the 15-chromatin state data for DLPFC tissue samples (E073) from the Roadmap Epigenomics Project[11]. Using combinations of histone modification marks, ChromHMM[12] was previously used to annotate segments of the genome with different chromatin states (repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions, and large-scale repressed and inactive domains), which were shown to vary across sex, tissue type, and developmental age[13]. Similarly, we tested enrichment of significant CpGs and DMRs in binding sites of transcription factors and chromatin proteins from the ENCODE project[14] and CODEX database[15] using the LOLA R package[16].

Finally, we performed pathway analysis by comparing the genes with significant DNA methylation changes in AD (identified in sex-stratified analysis) with the canonical pathways and biological process GO terms in MSigDB using GSEA analysis[17]. First, we linked each CpG and each pre-defined genomic region tested in DMR analysis (see Section "2. Single cohort analysis" above) to genes by annotating them using the GREAT (Genomic Regions Enrichment of Annotations Tool) software[18] (with default "Basal plus method"), which associates genomic regions to target genes. Next, we represented each gene by the smallest p-value if multiple CpGs or genomic regions are associated with them. To remove selection bias due to different numbers of CpGs or genomic regions associated with each gene (i.e., P-values from a gene with many CpGs or genomic regions linked to it are likely to be smaller than a gene with few linked CpGs or DMRs), we next fit a generalized additive model[19] using the R package mgcv: $Y_i \sim f(n.links_i)$ where Y_i is negative log (base 10) transformation of the P-value for gene i in the analysis of female samples (or male samples), $n.links_i$ is the number of CpGs or DMRs associated with gene i, and f is a penalized spline function. We assumed gamma distribution for Y_i, as under the null hypothesis of no association, Y_i follows the chi-square distribution (a special case of gamma distribution). The residuals from this model were estimated and used to generate a ranked gene list, which was then used as input for GSEA (in pre-ranked mode) to identify canonical pathways and gene ontology terms (MsigDB C2:CP and C5:BP collections of gene sets) enriched with significant methylation changes in female samples and male samples separately.

7. Integrative methylation – gene expression analysis

To systematically evaluate transcriptional changes near the sex-specific DNA methylation changes, we next performed integrative methylation – gene expression analysis using 529 (333 female and 196 male) samples from the ROSMAP study with matched DNA methylation and gene expression data. To this end, normalized FPKM (Fragments Per Kilobase of transcript per Million mapped reads) gene expression values for the ROSMAP study were downloaded from the AMP-AD Knowledge Portal (Synapse ID: syn3388564). First, we linked significant CpGs (or DMRs) to nearby genes using GREAT[18], which associates genomic regions to regulatory domains of genes. Next, we removed confounding effects in DNA methylation data by fitting the model methylation M value ~ neuron.proportion + batch + sample.plate + ageAtDeath and extracting residuals from this model; these are the *methylation residuals*. Similarly, we also removed potential confounding effects in RNA-seq data by fitting model log2(normalized FPKM values + 1) ~ ageAtDeath + markers for cell types. The last term, "markers for cell types," included multiple covariate variables to adjust for the multiple types of cells in the brain samples. More specifically, we estimated expression levels of genes that are specific for the five main cell types present in the CNS: ENO2 for neurons, GFAP for astrocytes, CD68 for microglia, OLIG2 for oligodendrocytes, and CD34 for endothelial cells, and included these as variables in the above linear regression model, as was done in a previous large study of AD samples[20]. The residuals extracted from this model are the *gene expression residuals*.

Finally, for each gene expression and CpG (or DMR) pair, we then tested the association between gene expression residuals and methylation residuals using a linear model: gene expression residuals ~ methylation residuals + Braak stage. For significant DMRs, this analysis was repeated, except that methylation M value was replaced with median methylation M value from multiple CpGs in the DMR.

8. Sex-specific mQTL analysis

To identify methylation quantitative trait loci (mQTLs) for the significant DMRs and CpGs, we tested associations between the methylation levels with nearby SNPs, using the ROSMAP study dataset, which had matched genotype data and DNA methylation data for 688 samples (434 females, 254 males). ROSMAP genotype data was downloaded from AMP-AD (syn3157325) and imputed to the Haplotype Reference Consortium r1.1 reference panel[21]. The male samples and female samples were analyzed separately.

To reduce the number of tests, we focused on identifying *cis* mQTLs located within 500kb from the start or end of the DMR (or position of the significant CpG)[22]. We additionally required SNPs to (1) have minor allele frequency of at least 1%, (2) be imputed with good certainty: information metric (info score) ≥ 0.4, and (3) be associated with AD case-control status (as determined by clinical consensus diagnosis of cognitive status), after adjusting for age, batch, and the first 3 PCs estimated from genotype data, at nominal P-value less than 0.05. We then fit the linear model methylation residual ~ SNP dosage + batch + PC1 + PC2 + PC3, where PC1, PC2, and PC3 are the first three PCs estimated from genotype data, to test the association between methylation residuals in CpGs and the imputed allele dosages for SNPs to identify mQTLs. The

analysis for DMRs was the same except that we replaced methylation residual with median (methylation residuals) of all CpGs located within the DMR.

References

1. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* **9**, 436-447 (2016).
2. Wang, T. *et al.* A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. *Epigenetics* **10**, 662-9 (2015).
3. Zhang, L. *et al.* Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. *Nat Commun* **11**, 6114 (2020).
4. Guintivano, J., Aryee, M.J. & Kaminsky, Z.A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290-302 (2013).
5. Gomez, L. *et al.* coMethDMR: accurate identification of co-methylated and differentially methylated regions in epigenome-wide association studies with continuous phenotypes. *Nucleic Acids Res* **47**, e98 (2019).
6. van Iterson, M., van Zwet, E.W., Consortium, B. & Heijmans, B.T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol* **18**, 19 (2017).
7. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**, 1539-58 (2002).
8. Rice K, Higgins J & T, L. A re-evaluation of fixed effect(s) meta-analysis. *J R Stat Soc A* **181**, 205-27 (2018).
9. Riley, R.D., Higgins, J.P. & Deeks, J.J. Interpretation of random effects meta-analyses. *BMJ* **342**, d549 (2011).
10. Van den Berge, K., Soneson, C., Robinson, M.D. & Clement, L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biol* **18**, 151 (2017).
11. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317-24 (2012).
12. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).
13. Yen, A. & Kellis, M. Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. *Nat Commun* **6**, 7973 (2015).
14. Davis, C.A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).
15. Sanchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res* **43**, D1117-23 (2015).
16. Sheffield, N.C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-9 (2016).
17. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
18. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
19. Wood, S.N. mgcv: GAMs and Generalized Ridge Regression for R. *R News* **1**, 20-25 (2001).
20. De Jager, P.L. *et al.* Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat Neurosci* **17**, 1156-63 (2014).
21. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).

22. Hannon, E. *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol* **17**, 176 (2016).