Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yuchen Yan Date

Cell-type specific alteration of
DNA methylation in Alzheimer's Disease

By

Yuchen Yan

Master of Science in Public Health

Biostatistics and Bioinformatics

Hao Wu, PhD
Thesis Advisor

Suprateek Kundu, PhD

Reader

Cell-type specific alteration of DNA methylation in Alzheimer's Disease

By

Yuchen Yan

B.S.
Dalian University of Technology
2017

Thesis Committee Chair: Hao Wu, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics and Bioinformatics
2019

Abstract

Alzheimer's disease (AD) is one of the chronic neurodegenerative disorders that causing great social burden. Hoping to identify novel biological signals in AD, we conduct in-depth analyses of brain DNA methylation data from ROS/MAP cohort, with consideration of cellular heterogeneity in brain tissues. We apply a reference-based method EpiDISH to estimate cell-type proportions, and a new method TOAST to detect cell-specific differential DNA methylation (csDM). The estimated portions show modest correlations with a number of clinical outcomes, including Braak stage, CERAD score, sex, overall amyloid level, age of death, and cognitive values. The csDM analysis does not find any cell-type specific differentially methylated loci with statistical significance after multiple testing correction. However, a more powerful joint test procedure identifies 1454 significant loci from the joint signals of glia and neuron. We systematically investigate the biological implication of the loci. From gene ontology (GO) enrichment analysis, we find that the transplanting mesenchymal stem cell (MSC) can be seen as potential way to stop AD, because these cells can express feature of the neural cell and have similarity with ependymal cells. It is inspiring because intracerebral transplantation of MSCs has been identified improvement in AD mice. This project provides a unique view to AD epigenetic research from cell-type specific analysis. Future studies could address the transplantation of MSCs method in AD area to validate new treatment and understand biological progress associated with AD, and to discover diagnostic biomarkers and therapeutic targets.

Acknowledgement

I would express great thanks to my thesis advisor Dr. Hao Wu, Associate Professor at our Department of Biostatistics and Bioinformatics at Emory University. Prof. Wu has been steering me in the right direction for writing paper with polishing layout and language in every sentence in thesis progress. I would also express great thanks to Dr. Ziyi Li, Postdoc at Emory University. Dr. Li patiently explained everything when I had problems. She gave me much help in the layout and content of my thesis. I would also express great thanks to my academic advisor Dr. Suprateek Kundu, for he gave me comments in detail to help me polish my paper. I would also thank my parents for giving me the chance to go to Emory University. Accomplishment and achievement would not have been possible without them. Thank you very much. ROS/MAP study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

Table of Contents

Introduction

Alzheimer's disease (AD) is one of the most prevalent chronic neurodegenerative disorders. It begins slowly with mild memory loss, and then worsens over time. The incidence of AD increases exponentially with age, and doubles every 5 years after the age of 65 (Kukull et al. 2002). During last several decades, AD has increasingly become a major global burden. Around the world, there are about 50 million people suffering from dementia (World Health Organization 2017), where AD consists of 60%-80% of the dementia (Alzheimer's association 2019; Deepali J. Mane 2018). The number of AD patients are estimated to be 75 million by mid 21st century, twice as many as the current number, assuming no effective treatment for AD (D. Selkoe, Mandelkow, and Holtzman 2012). Currently, the estimated average life expectancy after diagnosis is three to nine years (Deepali J. Mane 2018). AD treatments can temporarily slow down the dementia process, but no treatment can stop or cure AD. The cause of AD is still unclear, where the possible causes include genetic mutation, head injuries, depression, hypertension, obesity, mental exercise time, physical fitness level, and lifestyle (Deepali J. Mane 2018). For example, existing studies showed some AD cases are

associated with mutations on amyloid precursor protein (*APP)*, *PSEN1* and *PSEN2* genes (Gasparoni et al. 2018). However, these only contribute a small proportion of AD, where the major causes are still unknown (Gasparoni et al. 2018; Yokoyama, Rutledge, and Medici 2017). Worldwide efforts have been made to look for diagnostic biomarkers and therapeutic targets in order to design better treatment, delay the AD onset, and prevent AD from the start (Kukull et al. 2002).

Tremendous efforts have been spent to look for molecular mechanisms of AD. Large-scale population level studies have been conducted in genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS) (Allen et al. 2012; Barrachina and Ferrer 2009; Gasparoni et al. 2018). Genome-wide association studies (GWAS) detects associations between phenotypes of interest (such as human diseases) and genetic variations (Paul and Beck 2014). Similar to GWAS, epigenome-wide association studies (EWAS) can identify association between phenotype and epigenome changes at specific loci (Verma 2016; Zou et al. 2014).

In aging research, DNA methylation has gained intense interests because epigenome variations of DNA are heritable and can induce stable modifications in regulating gene

expression (Barrachina and Ferrer 2009; Paul and Beck 2014). In eukaryotes, DNA methylation mainly occurs at cytosine's that are followed by guanines (referred to as CpG site) (Feinberg et al. 2014). DNA methylation is closely related to many biological processes and human diseases. For example, aberrant DNA methylation is a hallmark of cancer (Barrachina and Ferrer 2009; Esteller 2002; Nelson 2007; Teschendorff et al. 2017). Therefore, people often look for in DNA methylation that are associated with disease, with hope to identify epigenetic biomarkers and therapeutic targets (Liu et al. 2013). In addition, recent findings have reported that aging and AD are associated with abnormal DNA methylation (Chouliaras et al. 2013; Gasparoni et al. 2018; Li et al. 2018).

There have been interests to identify modifications of DNA methylation in brain regions of AD cases (Yokoyama, Rutledge, and Medici 2017). An effective way for such task is the differential methylation (DM) analysis, where one conducts statistical tests on all CpG sites and identify the ones associated with outcome of interest, such as AD (Li et al. 2018). However, brain is a complex organ with many cell types, where different type of brain cells present highly heterogeneous functions and variable

genomic profiles (Zou et al. 2014). Traditional DM analysis and EWAS ignores the cell type mixture problem, which would lead to spurious associations (Li et al. 2018; Zou et al. 2014). The main goal of this work is to conduct in-depth re-analyses of existing DNA methylation data from AD brains, using several newly developed statistical methods to account for the cellular heterogeneity in brain tissues. With such consideration in the data analyses, we hope to identify novel biological signals.

Accounting for cellular heterogeneity in DM and EWAS has gain much interests lately, and several deconvolution methods dealing with cell-type specific proportions in complex tissues have been proposed. The deconvolution methods can be mainly divided into two categories: reference-based methods, which use DNA methylation reference profiles to conduct deconvolution based on regression models; and reference-free methods, which don't require the reference profiles and rely on some type of factor analysis to estimate proportions. Reference-based method are considered to be more accurate and stable than reference-free method, thus it is the safest option given reference DNA methylation profiles (Teschendorff et al. 2017). In this work, we will use a recently developed reference-based method EpiDISH (Epigenetic Dissection of

Intra-Sample-Heterogeneity). EpiDISH use robust partial correlations (RPC) method to estimate cell-type specific proportion. This method was reported to outperform competing reference-based methods in simulation and real data analyses. The R package is available on GitHub (https://github.com/sjczheng/EpiDISH) (Teschendorff et al. 2017).

After having estimated cell-type specific proportions, cell-type specific differential methylation (csDM) test then can be applied. We conduct csDM analysis using a newly developed statistical method, called TOAST (TOols for the Analysis of heterogeneouS Tissues) (Li et al. 2018). TOAST characterizes the data from mixed sample by a rigorous statistical framework and provide functionalities for flexible cell type specific test based on linear model framework. The R package is available on GitHub (https://github.com/ziyili20/TOAST).

In this project, we will focus on csDM analyses for AD on two types of brain cells: neuron and glia. Even though there are many cell types in brain, we are only able to find high quality reference methylomes for these two cell types. So we will focus on the results on these two cell types. We obtain data from GEO database with accession

number GES41826 as reference methylomes. Then EpiDISH package is applied to infer cell-type specific proportions of every subject of AD patients and controls. We then use TOAST package to conduct csDM test between AD patients and controls and compare the results with the ones from existing studies. We further conduct a number of analyses to interpret the biological meanings of the results.

Method

Description of ROS/MAP data

The DNA methylation data of AD patients and controls were generated on the samples collected from two prospective cohort studies at Rush University Medical Center: the Religious Orders Study (ROS) and the Memory and Aging Project (MAP) (De Jager et al. 2014). The ROS began in 1994 and ended in 2011, including 1168 persons who are above 53 years old (David A Bennett et al. 2012). The MAP started in 1997 and completed in 2011, including 1556 retire people who are above 53 years old (D.A. Bennett et al. 2012). All participants of the two studies were free from dementia when enrolled. The dataset used in this project is obtained from De Jager et al. (2014), who selected a random subset of older populations with 734 subjects. After extracting DNA

from the frozen dorsolateral prefrontal cortex of each subject, the methylation profiles were measured at 339,162 CpG sites using Illumina HumanMethylation450 beadset (De Jager et al. 2014). The DNA methylation data are transformed into beta values, ranging from 0 to 1. Thus, the input data is a beta value matrix with 339,162 rows (for CpG sites) and 734 columns (for subjects).

Since there's no diagnosis results for these patients, we choose to determine the AD status for all subjects based on their Braak stages. The Braak stage is a semi-quantitative measure of neurofibrillary tangles which have biological relation with tau protein (D. J. Selkoe 2012). The tau protein is abnormal in AD (Kolarova et al. 2012), and therefore, Braak stage can be an indicator of AD (Braak and Braak 1991). In our analysis, subjects with Braak stage of 0-3 are deemed as normal controls, and Braak stage of 4-6 are deemed AD patients. By these criteria, we have 368 normal control and 366 AD patients.

Reference DNA methylation data

The DNA methylation reference is constructed from the GEO dataset with accession number GES41826 (Kaminsky, Guintivano, and Aryee 2013). This study measures DNA methylation profiles of post mortem frontal cortex tissues from 58 subjects, half

of whom are major depression patients and the other half are matched controls, using Illumina 450k microarrays. We only use the data of cell-sorted neuron and glia from 29 controls. The data from these people are averaged to obtain the reference methylomes for neuron and glia.

Solving for proportions using EpiDISH

Given DNA methylation profiles of ROS/MAP, and brain reference methylomes for neuron and glia, we applied Bioconductor package EpiDISH to solve for cell-type specific (glia and neuron) proportions (Teschendorff et al. 2017). EpiDISH models a given DNA methylation profile as a linear combination of a known set of DNA methylation reference profiles:

$$Y = \sum_{k=1}^{K} {X_k}^T W_k + \epsilon$$

In the model, *Y* denote the DNA methylation profile from mixed tissue, which is the ROS/MAP data; *K* is the number of cell types, i.e. two in this application; X_k denotes the reference DNA methylation profiles of two cell types, which is constructed from cell-sorted glia and neuron data. The only unknown parameter in this model is W_K, which is the mixture proportions of two cell types. For each subject i, assuming

$\sum_{k=1}^{K} W_{ik} = 1$, EpiDISH uses robust partial correlation to estimate W_K (Teschendorff et al. 2017). After obtaining the initial estimate of W_K, EpiDISH normalize the proportion estimates by setting negative weights to zero and scaling all non-zero weights to make each row sums up to 1 (Houseman et al. 2012). We then obtain the mixing proportions of glia and neuron for each subject from ROS/MAP cohort.

We explore correlations between neuron proportions and other covariates, and perform statistical tests to evaluate the significance of the correlations. For categorical variables, we perform ANOVA analysis on the null hypothesis that neuron proportions have no significant differences across Braak stages, CERAD scores and sex group. For continuous variables, we fit linear regression model between neuron proportions and each continuous variable to test their correlation.

Cell-type specific DM (csDM) test using TOAST

With estimated mixture proportions of AD patients and controls, we detect neuron and glia specific DM signals using TOAST package. TOAST assumes personalized cell-type specific profile Π_{ik} for sample i and cell type k by model it as:

$$\mathrm{E}[\Pi_{ik}] = \mu_k + Z_i^T \beta_k$$

where μ_k is the baseline DNA methylation profile of cell type *k*, Z_i is a scaler or vector representing the subject's covariates (such as disease status, age, gender, etc.), and β_k is corresponding covariate effect(s). Following the previous notation, we denote the estimated mixture proportions by W_{ik}. Although the cell-type specific pure profile Π_{ik} is not directly observed, we can observe mixed signal Y_i for subject *i* and build connections between Y_i and Π_{ik} through a linear model:

$$\mathrm{E}[\mathrm{Y}_i; W_i] = \sum_k W_{ik}\,\mathrm{E}[\Pi_{ik}] = \sum_k (W_{ik}\,\mu_k + W_{ik} Z_i^T \beta_k).$$

In this model, Y_i, W_{ik} and Z_i are observed (or estimated), and μ_k and β_k are unknown parameters to be solved for. We can reformat this linear model to matrix form by denoting the observed data as $Y = (Y_1, Y_2 \ldots Y_N)^{\mathrm{T}}$, the design matrix V and the covariate vector β as

$$\mathrm{V} = \begin{bmatrix} W_{11} & W_{12} & \ldots & W_{1K} & W_{11}\cdot Z_1^T & W_{12}\cdot Z_1^T & \ldots & W_{1K}\cdot Z_1^T \\ W_{21} & W_{22} & \ldots & W_{2K} & W_{21}\cdot Z_2^T & W_{22}\cdot Z_2^T & \ldots & W_{2K}\cdot Z_2^T \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ W_{N1} & W_{N1} & \ldots & W_{NK} & W_{N1}\cdot Z_N^T & W_{N1}\cdot Z_N^T & \ldots & W_{NK}\cdot Z_N^T \end{bmatrix}$$

$$\beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

With estimated parameters, TOAST detects cell-type specific differential signals through testing linear combinations of the regression coefficients. Here we use the ROS/MAP DNA methylation data and the solved mixture proportion as inputs for TOAST to detect neuron- and glia-specific DM sites between AD patients and controls. Specifically, we test three null hypotheses: $\beta_1 = 0$ for glia-specific signals, $\beta_2 = 0$ for neuron-specific signals, and $\beta_1 = \beta_2 = 0$ for the joint signals from glia and neuron. Statistical results are corrected for multiple-testing by Benjamini-Hochberg False Discovery Rate (FDR). CpG sites with FDR less than 0.05 are deemed as cell-type specific DM. Then each differentially methylated CpG (DMC) is matched with associated gene using Illumina Infinium methylation 450k methylation microarrays of human genome version 19 (hg19) using R/Cran package *IlluminaHumanMethylation450kanno.ilmn12.hg19*.

Pathway and Gene Ontologies Analysis

We use all identified DMCs to conduct pathway analysis and gene ontology (GO) analyses using genes enrichment analysis tool *EnrichR* (McDermott et al. 2016) web server (http://amp.pharm.mssm.edu/Enrichr/). We set maximum entry as 1000 genes

and use hg19 as reference gene sets libraries. The outputs consist of three enrichment scores to represent the significance of overlapping input DMCs list and gene sets libraries. The first one is p-value of Fisher Exact test, a proportion test based on the assumption that each gene of DMC is binomially distributed, and the probability is independent for each gene within gene sets. The second one is a z-score using standard deviation from expected ranking in each gene set library. The third is a combined score, which integrates log p-values from Exact Fisher test and z-score of the deviation from expected rank. We choose combined score to represent overlapping because this score has the properties of both method and has best results compared with other score schemes (Meirelles et al. 2013). We use Kyoto Encyclopedia of Genes and Genomes (KEGG) cell signaling pathway database in 2016 to conduct pathway analysis and we use GO Biological ontological database in 2018 to conduct GO analysis (McDermott et al. 2016).

Results

Data Description

The reference DNA methylation data contains 480,492 CpGs with glia and neuron profiles from 29 subjects. DNA methylation data of ROS/MAP include 340,516 CpGs for 734 subjects. The overlaps between the two data are 339,162 CpGs. We summarized phenotype variables for 734 subjects in Table 1. Three categorical variables list in Table 1: Braak stage, CERAD score, and sex. Braak stage from 0 to 6, with number of subjects of 9, 63, 78, 218, 201, 158, and 7, respectively. CERAD score is a semi-quantitative measure based on neuritic plaque density for determining AD: value 1 means no AD, 2 means possible AD, 3 means probable AD, 4 means definite from the recommendation of Consortium to Establish a Registry for Alzheimer's Disease (CERAD) (Morris et al. 1988). Sex also makes difference with respect to risk of AD (Letenneur et al. 1999, 2012). Number of subjects for each CERAD score and both sex (female and male) for AD controls and patients are also shown. We also summarize six continuous variables for AD controls and patients: Global burden of AD pathology (gpath), overall amyloid level (amyloid), tangle density of neuronal neurofibrillary

(tangles), the age of death, education levels, and cognitive values. Global burden of AD pathology (gpath) is a quantitative summary value of AD pathologies (David A Bennett et al. 2018). Overall amyloid level(amyloid) can be a hallmark of AD when the amyloid level is abnormal (D. J. Selkoe 2012). Tangle density of neuronal neurofibrillary (tangles) is associated with aggression and depression in Alzheimer's patients (Lai et al. 2010). AD increases exponentially with age, and doubles every 5 years after 65 (Kukull et al. 2002). The status of AD also have association with education level (Letenneur et al. 1999, 2012). Cognitive test is an universal and effective way to show the neuropsychological memory disorders in clinical trials (Watson et al. 2014).

AD Status	**Controls**	**Patients**
Number of subjects	366	368
Braak Stage	0 = 9 / 1 = 63	4 = 201/ 5 = 158
	2 = 78 / 3= 218	6 = 7
CERAD score	1 = 44 / 2 = 99	1 = 179 / 2 = 145
	3 = 54 /4 = 61	3 = 13 / 4 = 29
Sex	F = 211 / M = 157	F = 255 / M = 111
Age (years) Mean ± SD (range)	86.02 ± 7.09 (65.99,106.50)	89.94 ± 5.60 (72.40 108.28)
Gpath Mean ± SD (range)	0.32 ± 0.35 (0.00, 1.65)	1.09 ± 0.60 (0.02, 3.18)
Amyloid Mean ± SD (range)	1.73 ±2.42 (0.00, 13.76)	5.24 ±3.92 (0.00, 19.11)
Tangles Mean ± SD (range)	2.09 ±2.18 (0.00, 14.82)	10.76 ±9.43 (0.02, 78.52)
Education (levels) Mean ± SD (range)	16.62 ±3.65 (7.00, 28.00)	16.17 ±3.53 (3.00, 25.00)

Cognitive value Mean ± SD (range)	0.00± 0.09 (-0.47, 0.14)	0.06 ±0.11 (-0.50, 0.18)

Table 1. Summary statistics for AD patients and controls from the ROS/MAP data.

Estimated cell Proportions

We select 1000 CpGs based on the most significant coefficient of variation as an index for marker genes among 339162 CpGs. After applied RPC method in EpiDISH using these index CpGs, we get the cell-type specific (glia and neuron) proportions. Figure 1 shows the portions of glia and neuron among AD and controls groups.

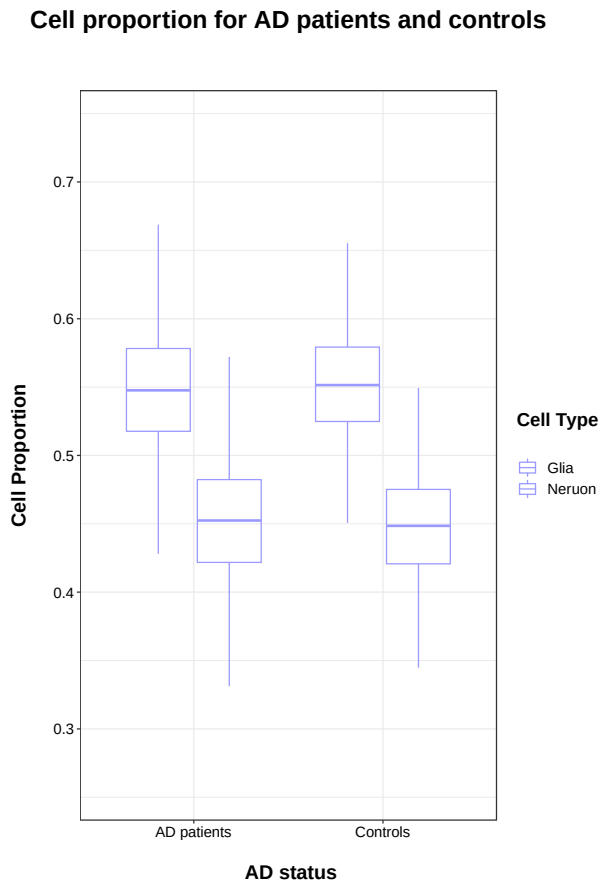

Figure 1. Proportions of glia and neuron for AD patient group and controls group

Cell type	AD Patients (SD)	Controls (SD)	Difference	p-value
Glia Proportion	0.549(0.0483)	0.552 (0.046)	0.003	0.378
Neuron Proportion	0.451(0.0483)	0.448 (0.046)		

Table. 2 Average proportions of glia and neuron among AD patient and control groups, with p-value under null hypothesis of no difference between two group.

The average proportions of glia and neuron among AD patient group are 0.549 (SD=0.0483) and 0.451 (SD=0.0483) respectively. The average proportions of glia and neuron among controls group are 0.552 (SD=0.046) and 0.448 (SD=0.046). Existing studies on the same brain regions (dorsolateral prefrontal cortex) show that the neuron cell number is 76.0-92.2 million, and non-neuron (glia) cell number is 74.8-94.4 million (von Bartheld, Bahney, and Herculano-Houzel 2016; Suzana 2014). Our estimated proportions match those numbers reasonably well. We apply two-sample t-test to compare the proportions of glia and neuron among AD patient and control groups. The result shows a non-significant difference (p=0.378), indicating that there's no obvious cell proportion changes between AD and control. To explore the correlation between neuron proportion and covariates in the data description section, we show the boxplot

for categorical variables and scatterplot with fitted LOESS curve for the continuous variable in Figure 2.

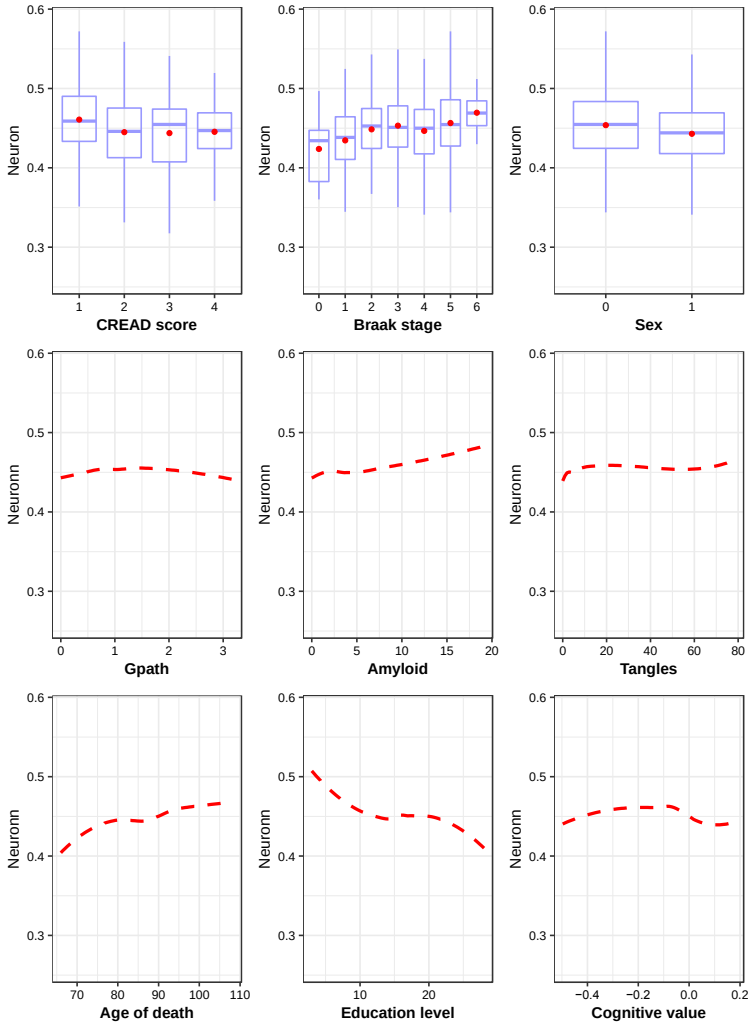

Figure 2. Proportion of neuron with Braak score, CREAD score, sex, gpath, amyloid, tangle, age of death, education, and cognitive values

From Figure 2, the proportions of neuron seem to have positive correlation with Braak stage, amyloid level, and age of death visually. Therefore, we perform ANOVA test under null hypothesis that neuron proportions have no significant differences across Braak stages, CERAD scores and sex group; and we conduct a linear regression analysis between neuron proportions and each continuous variable. The results show that there are indeed significant differences across the Braak stages ($p = 0.00361$), across the CERAD scores ($p = 0.00155$), and between females and males ($p = 0.00203$). The regression model suggest that neuron proportion have significant association with amyloid (coefficient = 0.0016, $p = 0.00044$), age of death (coefficient = 0.00099, $p =$ 9.4e-05), and cognitive value (coefficient = -0.059, $p = 0.00028$). In addition, the test results for other variables are: gpath (coefficient = 0.0049, $p=0.0679$), tangles (coefficient = 0.00040, $p=0.0575$), education level (coefficient = -0.00084, $p=0.0774$). These associations are weak, but still marginally significance. There have been some reports on the neuron proportion change with covariates, for example, the neuron proportion in brain have a positive association with age (Soreq et al. 2017). Our findings

suggest that the neuron proportions have weak, but statistically significant correlation with many AD-related covariates.

DMCs of glia, neuron, and joint analysis

For 339,162 input CpGs, we apply TOAST to look for glia- and neuron-specific DM between AD and control. The results are disappointing that none of the CpG show cell type specific significant changes after multiple testing correction. However, when we apply a joint test ($\beta_1 = \beta_2 = 0$ for testing either glia or neuron has difference between AD and control), we find 1454 significant DMCs. We also conduct DM calling without consideration of cell mixtures and found 1453 DMCs. Compare these two lists, we find that there are 1232 overlapping DMCs. We show the top 10 DMCs defined at lowest 10 FDR for joint-DMCs in Table 3.

CpGs	Chromosome	Position	Gene	P value	FDR
cg05066959	chr8	41519308	ANK1; MIR486	9.13E-17	3.10E-11
cg03169557	chr16	89598950	SPG7	2.54E-15	4.31E-10
cg11823178	chr8	41519399	ANK1; MIR486	6.90E-15	6.16E-10
cg26102082	chr17	47590272	NGFR	8.04E-15	6.16E-10
cg25018458	chr17	980014	ABR	1.01E-14	6.16E-10
cg13076843	chr17	74475294	RHBDF2	1.09E-14	6.16E-10
cg05810363	chr17	74475270	RHBDF2	1.60E-14	7.73E-10
cg22883290	chr2	127800646		2.80E-14	1.19E-09
cg16588649	chr19	3463241	NFIC	2.06E-13	7.78E-09

cg05066959	chr17	74475355	RHBDF2	2.88E-13	9.77E-09

Table. 3 Top 10 DMCs of joint signal defined at lowest 10 FDR.

Pathway Analysis and Go Analysis

Using *EnrichR*, we conduct KEGG pathway analysis for DMCs of the joint signal. However, there is no significant association between AD pathway and DMCs of joint signal (rank=113, p=combined scores=-0.48). We then conduct an enriched GO analysis. Five enriched GO terms for top 2 enriched GO progress with three amyloid-beta Go progress for the joint signal are summarized in Table 4.

Rank	Description	GO number	P-value	Adjusted p-value	Combined score
1	Positive regulation of mesenchymal cell proliferation	GO:0002053	0.00001053	0.03221	30.20
2	Regulation of mesenchymal cell proliferation	GO:0010464	0.0002751	0.1402	21.68
15	Negative regulation of amyloid-beta formation	GO:1902430	0.01126	0.5867	13.74
307	Regulation of amyloid-beta formation	GO:1902003	0.06534	0.7165	4.53
819	Positive regulation of amyloid-beta clearance	GO:1900223	0.2999	0.8644	1.94

Table 4. Top 2 and 3 Amyloid-beta enriched GO terms for joint signal.

Comparison with existing results

We compare results with those from a published paper on ROS/MAP DNA methylation data (De Jager et al. 2014). For De Jager et al. study, the paper reported 71 associated CpGs from 415,848 input CpGs from ROS/MAP of 708 subjects. Out of their 71 identified DMCs, 64 are included in our input data, and among them, 46 DMCs can match our results. We compare the ranking differences in Table 5, and find reasonably good correspondence.

Original Ranking	Ranking in Joint Signal	CpG	chr	Position(bp)	Gene
1	110	cg11724984	chr12	121890864	KDM2B
2	13	cg23968456	chr10	73521631	CDH23
3	30	cg15821544	chr1	43473840	
4	127	cg16733298	chr16	19127132	ITPRIPL2
5	24	cg22962123	chr7	27153605	HOXA3
6	6	cg13076843	chr17	74475294	RHBDF2
7	34	cg25594100	chr7	4786943	FOXK1
8	11	cg19803550	chr17	1637391	WDR81
9	2	cg03169557	chr16	89598950	SPG7; SPG7
10	1	cg05066959	chr8	41519308	ANK1; MIR486
11	7	cg05810363	chr17	74475270	RHBDF2
12	29	cg07012687	chr17	80195180	SLC16A3
13	732	cg21207436	chr14	74815316	C14orf115
14	38	cg21806242	chr11	72532891	ATG16L2
15	3	cg11823178	chr8	41519399	ANK1; MIR486
16	10	cg12163800	chr17	74475355	RHBDF2
17	113	cg17474422	chr1	36039866	TFAP2E

19	92	cg22904711	chr19	44278628	KCNN4
21	525	cg18556455	chr2	45178474	
22	279	cg05731218	chr2	216769199	
23	17	cg17693222	chr8	42033472	PLAT
24	185	cg12307200	chr3	188664632	
25	846	cg19007269	chr10	105420501	SH3PXD2A
26	148	cg15645660	chr1	55247356	TTC22
27	8	cg22883290	chr2	127800646	
28	253	cg14074251	chr2	220299116	SPEG
30	445	cg09448088	chr13	113635690	MCF2L
31	224	cg02308560	chr19	1071176	HMHA1
32	1264	cg20733077	chr13	50700845	DLEU2
33	853	cg13639901	chr7	155556590	RBM33
34	22	cg20618448	chr19	49962324	ALDH16A1
36	67	cg07883124	chr13	113634042	MCF2L
37	372	cg24231804	chr15	67316861	
38	1271	cg12877335	chr12	94539319	
39	151	cg22941668	chr5	148810180	MIR145; LOC728264
41	818	cg14430943	chr7	155556652	RBM33
42	370	cg24676346	chr6	41377288	
44	89	cg11652496	chr15	77324526	PSTPIP1
45	575	cg08737189	chr7	131223417	PODXL
50	415	cg15348679	chr21	45626491	
52	45	cg04157161	chr17	7906847	GUCY2D
55	489	cg22385702	chr2	45175881	
56	233	cg27443779	chr11	14664793	PDE3B; PSMA1
58	719	cg06742628	chr5	16886424	MYO10
60	1365	cg05322931	chr17	840950	NXN
62	1165	cg06753513	chr17	3977385	ZZEF1

Table 5. Ranking comparisons with paper (De Jager et al. 2014).

This difference may cause by different study methods and different thresholds. The detection is divided into three stages in De Jager et al. study. After detecting AD-associated DMCs, they replicated those DMCs in other sets of subjects to produce mRNA and then validated DMCs based on the mRNA. Also, they use Bonferroni correction under $P<0.05$ to detect significance, while we use FDR which has more power than Bonferroni correction. Therefore, we get more CpGs than the results of studies for our one stage design and FDR to represent Type I error.

Discussion

There are certain limitations for our study on the Illumina platform, such as the array's inability to distinguish DNA methylation and DNA hydromethylation. The cell proportions of AD patients and controls also have no significant difference. However, our result shows that neuron proportion and age of death have a significant positive linear association with a coefficient of 0.00099. Glia proportions decrease with age can correspond to the hypothesis that neurodegenerative disorders are caused by malfunctions of glia in the brain (Kaminsky, 2016). The brain has been identified as having neural stem cells, called ependymal cells (Johansson et al. 1999). Due to the limitation of the cell type for our reference data, we would speculate that if these glial

cells were ependymal stem cells and if the proportion change is mainly due to malfunctioning or other biological progress, ependymal cells would be an exciting topic concerning AD. Further study could separate cells into more cell types to detect the ependymal cells changes for AD.

We did not find any glia- or neuron- specific DMC, but we find 1454 joint-DMCs and shows a good match for another method and a similar study (Feinberg et al. 2014; De Jager et al. 2014). From joint-DMCs, the gene *ANK1, RHBDF2* have been identified having an association in AD pathway (De Jager et al. 2014). These two genes have essential functions in modulating the initiation of microglia and infiltrating macrophages (De Jager et al. 2014; Mastroeni et al. 2017). From ontology process, GO term ranked top which derived from five genes *TBX1, SHH, CHRD, LRP5, PDGFA, VEGFA*, is a process of activating mesenchymal cell proliferation. Mesenchymal stem cells (MSCs) are bone marrow-derived stem cell in adult, and MSCs can self-renewal and maintain their multipotency. Undifferentiated mesenchymal cells have been shown to express neural feature (Petersen et al. 2005), but the doubt exist of whether the neurons derived from MSC will be functional. The relation of MSCs and AD have been

identified in mice, and it shows significant improvement in AD mice for reducing amyloid depositions and increasing memory deficits with intracerebral transplantation of MSCs (Bae et al. 2009). We speculate that transplanted MSCs may show some similarity with ependymal cells under brain condition and in this way, MSCs can improve memory deficits in AD mice. Future studies can validate this method in mice and evaluate the method of intracerebral transplantation of MSCs, which could be a potential way to stop AD for human in the future.

Conclusion

This project provides a new view to study the epigenetic modification in AD, through cell type specific DM analysis. This cell-type specific idea in the differential analysis could provide insights for diagnostic biomarkers and therapeutic targets. Future studies could address cell-type specific idea in AD area to validate new treatment and discover more biological progress associated with AD.

Reference

Allen, Mariet et al. 2012. “Novel Late-Onset Alzheimer Disease Loci Variants Associate with Brain Gene Expression.” *Neurology* 79(3): 221–28.

Alzheimer’s association. 2019. “What Is Alzheimer ’ S ?”

Barrachina, Marta, and Isidre Ferrer. 2009. “DNA Methylation of Alzheimer Disease and Tauopathy-Related Genes in Postmortem Brain.” *Journal of Neuropathology and Experimental Neurology* 68(8): 880–91.

Bennett, D.A. et al. 2012. “Overview and Findings from the Rush Memory and Aging Project.” *Current Alzheimer Research* 9(6): 646–63.

Bennett, David A et al. 2018. “Religious Orders Study and Rush Memory and Aging Project.” *Journal of Alzheimer’s disease : JAD* 64(s1): S161–89. https://www.ncbi.nlm.nih.gov/pubmed/29865057.

Bennett, David A, Julie A Schneider, Zoe Arvanitakis, and Robert S Wilson. 2012. “Overview and Findings from the Religious Orders Study.” *Current Alzheimer research* 9(6): 628–45.

Braak, H, and E Braak. 1991. “Neuropathological Stageing of Alzheimer-Related Changes.” *Acta neuropathologica* 82(4): 239–59.

Chouliaras, Leonidas et al. 2013. “Consistent Decrease in Global DNA Methylation and Hydroxymethylation in the Hippocampus of Alzheimer’s Disease Patients.” *Neurobiology of Aging* 34(9): 2091–99. http://dx.doi.org/10.1016/j.neurobiolaging.2013.02.021.

Deepali J. Mane, Dr. Mrs. Vanita Kanase and Mr. Imtiyaz Ansari. 2018. “AN OVERVIEW OF TREATMENT FOR ALZHEIMER’S DISEASE.” 5(6): 170–75.

Esteller, Manel. 2002. “CpG Island Hypermethylation and Tumor Suppressor Genes: A Booming Present, a Brighter Future.” *Oncogene* 21(35 REV. ISS. 3): 5427–40.

Feinberg, Andrew P. et al. 2014. “Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays.” *Bioinformatics* 30(10): 1363–69.

Gasparoni, Gilles et al. 2018. “DNA Methylation Analysis on Purified Neurons and Glia Dissects Age and Alzheimer’s Disease-Specific Changes in the Human Cortex.” *Epigenetics and Chromatin* 11(1).

Houseman, Eugene Andres et al. 2012. “DNA Methylation Arrays as Surrogate Measures of Cell Mixture Distribution.” *BMC bioinformatics* 13(1): 86. http://www.biomedcentral.com/1471-2105/13/86%5Cnhttp://www.bloodjournal.org/lookup/doi/10.1182/blood-2008-06-162958.

De Jager, Philip L. et al. 2014. “Alzheimer’s Disease: Early Alterations in Brain DNA Methylation at ANK1, BIN1, RHBDF2 and Other Loci.” *Nature Neuroscience* 17(9): 1156–63. http://dx.doi.org/10.1038/nn.3786.

Kaminsky, Jerry Guintivano, and Martin J Aryee. 2013. “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(3): 290–302. http://f1000.com/717997792.

Kolarova, Michala et al. 2012. “Structure and Pathology of Tau Protein in Alzheimer Disease.” 2012.

Kukull, Walter A. et al. 2002. “Dementia and Alzheimer Disease Incidence: A Prospective Cohort Study.” *Archives of Neurology* 59(11): 1737–46.

Lai, Mitchell K P, Christopher P Chen, Tony Hope, and Margaret M Esiri. 2010. “Hippocampal Neurofibrillary Tangle Changes and Aggressive Behaviour in Dementia.” *Neuroreport* 21(17): 1111–15.

Letenneur, L. et al. 2012. “Education and Risk for Alzheimer’s Disease: Sex Makes a Difference EURODEM Pooled Analyses.” *American Journal of Epidemiology* 151(11): 1064–71.

Letenneur, L et al. 1999. “Are Sex and Educational Level Independent Predictors of Dementia and Alzheimer’s Disease? Incidence Data from the PAQUID Project.” *Journal of Neurology, Neurosurgery & Psychiatry* 66(2): 177 LP-183. http://jnnp.bmj.com/content/66/2/177.abstract.

Li, Li et al. 2018. “Dissecting Differential Signals in High-Throughput Data from Complex Tissues.” *bioRxiv Bioinformatics*. http://biorxiv.org/cgi/content/short/402354v1.

Liu, Yun et al. 2013. “Epigenome-Wide Association Data Implicate DNA Methylation as an Intermediary of Genetic Risk in Rheumatoid Arthritis.” *Nature Biotechnology* 31(2): 142–47.

McDermott, Michael G. et al. 2016. “Enrichr: A Comprehensive Gene Set

Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research* 44(W1): W90–97.

Meirelles, Gabriela et al. 2013. "Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool." *BMC Bioinformatics* 14(1): 128.

Morris, J C et al. 1988. "Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Clinical and Neuropsychological Assessment of Alzheimer's Disease." *Psychopharmacology bulletin* 24(4): 641–52.

Nelson, William, G. 2007. "Abnormal DNA Methylation, Epigenetics, and Prostate Cancer." *Frontiers in Bioscience* 12(8–12): 4254.

Paul, Dirk S., and Stephan Beck. 2014. "Advances in Epigenome-Wide Association Studies for Common Diseases." *Trends in Molecular Medicine* 20(10): 541–43. http://dx.doi.org/10.1016/j.molmed.2014.07.002.

Selkoe, Dennis J. 2012. "Preventing Alzheimer's Disease." 337: 1488–92. www.sciencemag.org/special/prevention.

Selkoe, Dennis, Eckhard Mandelkow, and David Holtzman. 2012. "Deciphering Alzheimer Disease." *Cold Spring Harbor Perspectives in Medicine* 2(1): 1–8.

Soreq, Lilach et al. 2017. "Major Shifts in Glial Regional Identity Are a Transcriptional Hallmark of Human Brain Aging." *Cell Reports* 18(2): 557–70. http://dx.doi.org/10.1016/j.celrep.2016.12.011.

Teschendorff, Andrew E., Charles E. Breeze, Shijie C. Zheng, and Stephan Beck. 2017. "A Comparison of Reference-Based Algorithms for Correcting Cell-Type Heterogeneity in Epigenome-Wide Association Studies." *BMC Bioinformatics* 18(1): 1–14.

Verma, Mukesh. 2016. "Genome-Wide Association Studies and Epigenome-Wide Association Studies Go Together in Cancer Control." *Future Oncology* 12(13): 1645–64.

Watson, Jennifer L, Laurie Ryan, Nina Silverberg, and Marie A Bernard. 2014. "Obstacles And Opportunities In Alzheimer's Clinical Trial Recruitment." 33(4): 574–79.

World Health Organization. 2017. "Dementia." : 1–8. https://www.who.int/en/news-room/fact-sheets/detail/dementia.

Yokoyama, Amy S., John C. Rutledge, and Valentina Medici. 2017. "DNA Methylation Alterations in Alzheimer's Disease." *Environmental Epigenetics*

3(2): 1–11.
http://academic.oup.com/eep/article/doi/10.1093/eep/dvx008/3866541.

Zou, James et al. 2014. “Epigenome-Wide Association Studies without the Need for Cell-Type Composition.” *Nature Methods* 11(3): 309–11.