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:

_____ _____

Lidan Zhang Date

Applying Genome-Wide Association Study to Analyze Novel Variants That Potentially Relate to Alzheimer's Disease

By

Lidan Zhang

Master of Public Health

Department of Biostatistics and Bioinformatics

_____

Zhaohui (Steve) Qin, Ph.D.
Thesis Advisor

_____

Thomas Wingo, M.D.
Reader

Applying Genome-Wide Association Study to Analyze Novel Variants That Potentially Relate to Alzheimer's Disease

By

Lidan Zhang

B.S.

Georgia State University

2018

Thesis Committee Chair: Zhaohui (Steve) Qin, Ph.D.

Reader: Thomas Wingo, M.D.

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 Public Health
in Biostatistics
2020

Abstract

Applying Genome-Wide Association Study to Analyze Novel Variants That Potentially Relate to Alzheimer's Disease

By Lidan Zhang

Alzheimer's disease is a progressive and irreversible brain disease that is a common cause of dementia. Alzheimer's patients will gradually lose their ability to think independently, remember, make decisions, and function independently. Alzheimer's disease is the sixth leading cause of death in the United States and there is no cure. In this study, we used datasets from ROSMAP, a project related to aging and dementia, to conduct Genome-Wide Association Study (GWAS). GWAS is an approach to relate gene data with specific diseases or traits. We used the gene expression data of 640 individuals with Alzheimer's disease and matched with 1608 GWAS SNPs. A differential expression analysis was conducted over the expression data for the GWAS SNPs. A gene set enrichment analysis (GSEA) was followed to determine which pathways show significance and are related to the cause of Alzheimer's disease.

Keyword: Alzheimer's disease, genome-wide association study, gene set enrichment analysis, SNP, differential expression analysis, pathway

Applying Genome-Wide Association Study to Analyze Novel Variants That Potentially Relate to Alzheimer's Disease

By

Lidan Zhang

B.S.

Georgia State University

2018

Thesis Committee Chair: Zhaohui (Steve) Qin, Ph.D.

Reader: Thomas Wingo, M.D.

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 Public Health
in Biostatistics
2020

Acknowledgement

I would like to first thank the Department of Biostatistics and Bioinformatics for the support and help on both academy and daily life, especially during this hard period of quarantine and learning remotely. I would like to thank Dr. Steve Qin, my thesis advisor who has advised and guided me to complete this thesis paper and provided so much help. I am also grateful that Prof. Thomas Wingo from Departments of Neurology and Human Genetics, School of Medicine, agreed to be my reader and provided his precious comments. I am also appreciated that my family and my friends kept motivating me and took good care of me. They were always by my side when I was stressful and anxious. So much thanks to my parents living oversea who sent me encouraging words everyday even there were 12-hour time difference. Without all the support, I would not be able to accomplish the thesis.

Table of Contents

Introduction

Alzheimer's disease is a progressive and irreversible brain disease that is a common cause of dementia [1]. In normal human brains, neurons continuously communicate with each other through electrical charges that travel through axons, causing chemicals to be released through synapses. However, among people with Alzheimer's disease, toxins such as beta-amyloid and tau accumulate and spread in the brain, cutting off connections between neurons [1]. The vascular problem may be another reason that the brain cannot provide enough blood and nutrition. As neurons die, the brain shrinks, beginning from the hippocampus, which is an important part of brain learning and memory [1][2]. Alzheimer patients will gradually lose their ability to think independently, remember, make decisions, and function independently [1]. According to the CDC, Alzheimer's disease is the sixth leading cause of death in the United States [3]. There is no cure for Alzheimer's disease in the current stage. Treatment options may include drugs that reduce toxins that block the communication between neurons, treatments aiming to the vascular system, care plans, interventions such as exercise, and social engagement to improving brain health [3][4].

The cause of Alzheimer's disease is permanent gene mutations which are reversion, transformation, addition, or deletion of genes [5]. Genetic risk factor refers to genetic variants that increase the risk of disease but doesn't directly cause the disease. Sometimes a gene does not affect only one trait, which is called pleiotropic gene. Pleiotropy implies a mapping from one thing at the genetic level to multiple things at a phenotypic level, which means that a gene can result in two or more different phenotypic traits [6][7][8]. The ε4 allele of the apolipoprotein E (APOE) is the best-known genetic risk factor for Alzheimer's [1][9]. The risk of Alzheimer's elevated by the APOEε4 allele is 3 times in

heterozygotes and 15 times in homozygotes. It's been proven that 40%-80% of Alzheimer patients have at least one APOEε4 allele. Other genetic risk factors that been proven to be related to Alzheimer's diseases include: CASS4, CELF1, FERMT2, HLA-DRB5, INPP5D, MEF2C, NME8, PTK2B, SORL1, ZCWPW1, SlC24A4, CLU, PICALM, CR1, BIN1, MS4A, ABCA7, EPHA1, and CD2AP [6][9][10].

To better understand Alzheimer's disease from a genetic perspective, genome-wide association study (GWAS) is applied to look for variants that could potentially cause Alzheimer's disease. GWAS can identify new genetic mutations in Alzheimer's disease based on genetic pathways and specific disease processes [11][12]. Previous studies have shown that tau and tau phosphorylated at threonine 181 (ptau) in Alzheimer's patients are higher than in non-demented elderly [12]. Genetic mutations that have been shown to increase the risk of Alzheimer's disease alter the levels of tau, including pathogenic mutations in APP, PSEN1 and PSEN2, and common mutations in APOE [1][9][13]. Numbers of genes together consist of pathways which are regulatory networks. A gene can be involved in different pathways and play different roles. Gene set enrichment analysis (GSEA) is a method to classify genes into pathways. The GSEA uses predefined datasets and analyze genes falling into which pathways [14]. Using GSEA helps to find potential pathways associated with Alzheimer's.

The dataset using in this study is from the Religious Orders Study and the Rush Memory and Aging Project (ROSMAP). The ROSMAP cohort is designed for studies related to aging and dementia. It's a GWAS dataset for meta-analysis [15]. Previous study of cognitive testing based on this dataset has found that APOE is strongly associated with cognitive decline [16]. In this study, ROSMAP cohort

is investigated with 55889 genes and 642 individuals assembled and a genome-wide association study (GWAS) is conducted to detect novel variants that may potentially link to Alzheimer's disease.

Methods and Materials

Data

The data ROSMAP is a GWAS dataset from NIAGADS. It is composed of two studies: The Memory and Aging Project (MAP) and the Religious Orders Study (ROS). Both the studies are cohort studies of aging and dementia that include organ donation at death. Together, more than 2,700 persons have participated into clinical evaluation and have agreed to brain donation at death. ROSMAP is a GWAS dataset and is used in ADGC meta-analysis linked to dementia and neurodegeneration. It is one of the five cohorts assembled by ADGC to replicate genetic variant findings in LOAD [15].

The expression data contains the expression level data in FPKM for each individual and each gene. Patients' IDs are in columns and the ensembl gene IDs are in rows. Ensemble gene ID in this dataset is consisted of "ENS", "G" for gene, and unique numeric identifier for objects [17]. In this dataset, the expression level data of 55889 genes are recorded for 640 individuals. The genotype data contained information of 382 individuals including IDs, sex, and genotypes. The formats of genotype data are in plink (.bed, .bim and .fam) format (will be introduced later). There was a list of SNPs that we wanted to conduct GWAS and the list contained 1608 SNPs.

Data Processing

To extract SNP genotype information, we used PLINK [18][19]. The version we used was v1.9 for MS-DOS. The genotype data were originally in .bed, .bim, and .fam format since this set of format had a smaller size and were easy for transferring. Genotype information of GWAS SNP list was extracted. The output data were in .ped and .map format for further analysis.

For the following steps of analysis, we used R Studio. The expression data was compared with the filtered .ped PLINK file from last step. Individuals with the same IDs between expression level data and genotype data were kept for analysis. Information of sex, paternal and maternal ID, and phenotype was deleted since only individual ID and genotypes were used. SNPs with minor allele frequency (MAF) smaller than 0.1 were removed. MAF is the frequency in which the second most common allele occurs in a given population. Small MAF indicates a small frequency of the minor allele mutation, which will give a bias to our analysis. Then, the genotypes of each SNP were divided into two groups: homozygous major allele versus heterozygous and homozygous minor allele. Heterozygous and homozygous minor allele were considered as one group to assess what pathways would be influenced by minor allele mutation. SNPs with unknown genotypes were removed so that every individual and every SNP had known genotypes.

The expression data had gene ID in ensemble gene ID format. In order to conduct the following gene set enrichment analysis, gene IDs in different format were obtained by the package "biomaRt". The biomaRt database being used to convert gene ID was "hsapiens_gene_ensembl" containing genes of homo sapiens, since our targets were Alzheimer's disease patients. The list contained information including ensembl gene ID, entrez gene ID, HGNC symbol, external gene name,

description, chromosome name, and strand. The expression data was transformed to log type, which was $log_2(data + 0.1)$, to make the data less skewed.

Differential Expression (DE) Analysis

To conduct differential expression (DE) analysis, the package we used was "limma" [20]. The genotype was first made into group lists indicating which group the individuals were in for each SNP. Then the group lists were converted into group matrix with the first column representing group 1 and the second column representing group 2. In those two columns, "1" indicated the individual was in the group and "0" indicated not in the group. The row names were individual IDs and each SNP had a corresponding group matrix. Contrast matrix were also made by limma function to express contrasts between a set of parameters as a numeric matrix. The parameters were the group variables which were the coefficients of the linear models that would be fit later. The contrast matrix specified which comparisons between the coefficients are to be extracted from the fit. When the preparations were done, the data was fit into a linear model. The function that was used was called "lmFit" in "limma" package. It fit our data into a linear model for each gene. The log-transformed expression data was response variable and the group was categorical variable. The next step was to compute estimated coefficients and standard errors using the function "contrasts.fit". The idea of this function was to fit a full-rank model using lmFit, then use contrasts.fit to obtain coefficients and standard errors for any number of contrasts of the coefficients of the original model. Then the function "eBayes" was used to compute moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes moderation of the standard errors towards a common value with no trend default. It produced an object containing moderated t-statistics, p-values corresponding to the t-statistics, log-odds of differential expression, and other information. After the lists were made, tables of the top-ranked

genes from the linear model fit were extracted with NA values removed. The extracted tables were the results for differential expression analysis and each SNP had a corresponding table.

Gene Set Enrichment Analysis (GSEA)

For gene set enrichment analysis (GSEA), the R packages we used were: "fgsea" and "gageData" [21][22]. The gene IDs were in ensembl annotation and we obtained a list of gene information by "biomaRt" package. In order to conduct GSEA by fgsea package, the ensembl gene IDs were converted into external gene names. However, the pathway dataset used for this GSEA step was "kegg.sets.hs" from gagaData with entrez gene ID, so the format of gene ID in the pathway dataset was also converted into gene names in order to match with the DE results. Since the GSEA was conducted by the p-values generated by the DE analysis, vectors of p-values named by gene names were extracted for each SNP and the vectors were ranked in the increasing order. Then a GSEA procedure was done by the function "fgsea" using the pathway dataset with gene names and the p-value vectors, with 1000 numbers of permutation. Tables with GSEA results were generated with each row of tested pathways and columns of enrichment p-values, BH-adjusted p-values, enrichment scores, normalized enrichment scores, number of times a random gene set had a more extreme enrichment score value, size of the pathway after removing genes not present in 'names(stats)', and indexes of leading edge genes that drive the enrichment. Each SNP had a corresponding table of tested pathways.

Pathway Visualization

To visualize the relation between the pathways and SNPs, we did heatmaps for direct visualization. A matrix containing the GSEA p-values was generated with rows of SNPs and columns of tested pathways. The NA values were removed. To do heatmaps, the function "heatmap.2" from the

"gplots" package [23] was used with the setting of no trace, and color key and histogram. The color range was dark blue indicating small values, white indicating medium values, and dark red indicating large values. Enrichment plots were plotted for significant pathways through the function "plotEnrichment" in the "fgsea" package.

Results

The original expression data contained 55889 genes and 640 individuals. There were 282 individuals having matched IDs in the genotype data. Since the expression data were kept only for genes with corresponding external gene names, the number of remaining genes were 19196. The number of SNPs remained for DE analysis and GSEA was 345 out of 160.

Figure 1 is the heatmap of ranked datasets for gene set enrichment analysis. The heatmap has columns of tested pathways and rows of SNPs with p-values of differential analysis results in the table. Dark blue indicates small values, white indicates medium values, and dark red indicates large values. As the heatmaps were generated using p-values, the left side was p-value close to 0, and the right side was p-value close to 1. The smaller the p-value was, the more it was close to dark blue, and vice versa, the bigger p-value was, the more it was close to dark red. There are 18 pathways that somehow show significance across SNPs. In the 18 pathways, pathway "hsa04976 Bile secretion", "hsa04150 mTOR signaling pathway", "hsa04722 Neurotrophin signaling pathway", and "hsa00620 Pyruvate metabolism" show significance in all 345 SNPs, and pathway "hsa04142 Lysosome", "hsa00785 Lipoic acid metabolism", and "hsa04972 Pancreatic secretion" also show significance in

more than 95% of SNPs. Table 1 gives numbers of SNPs that pathways show significance in and the rate.

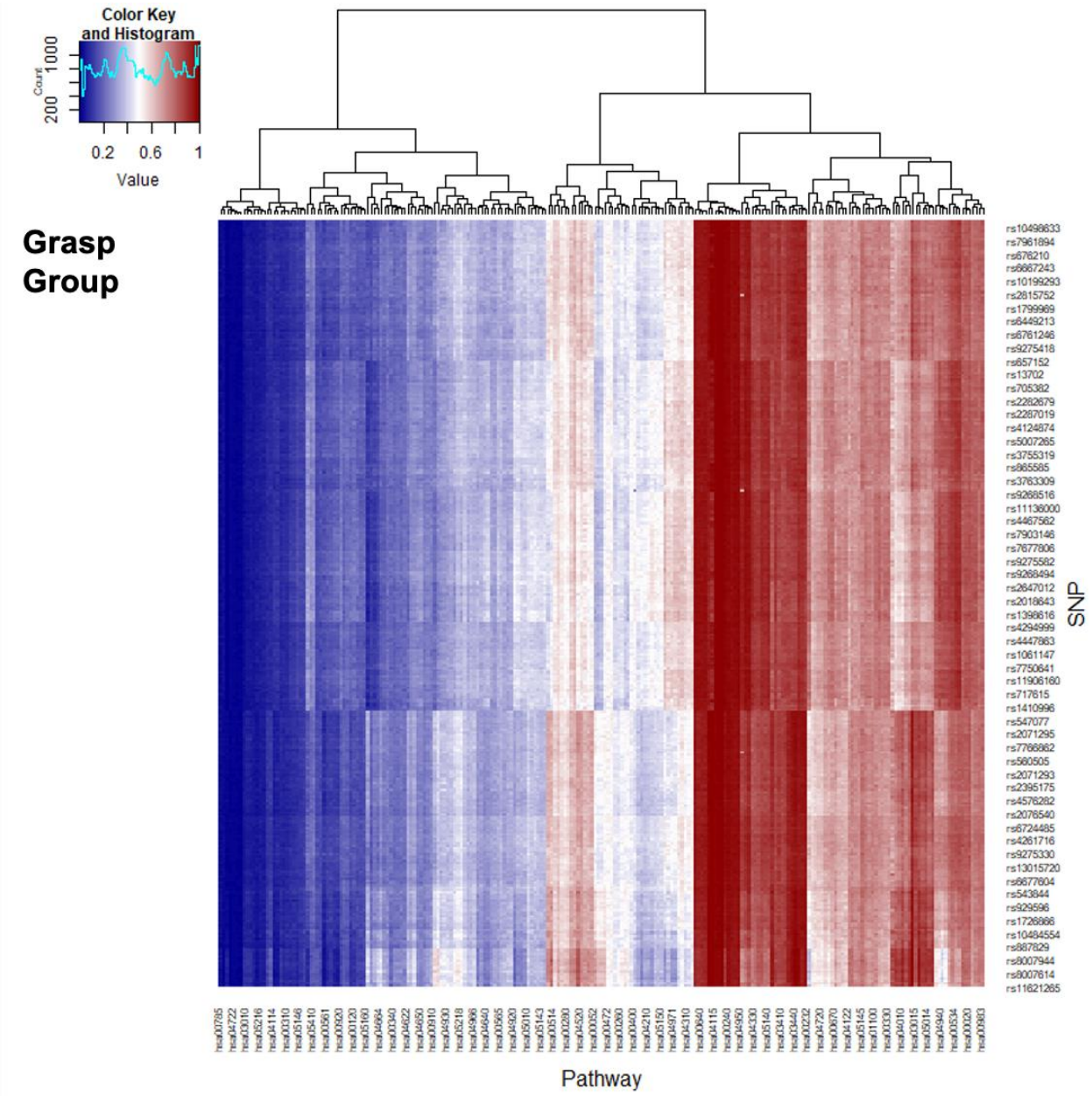


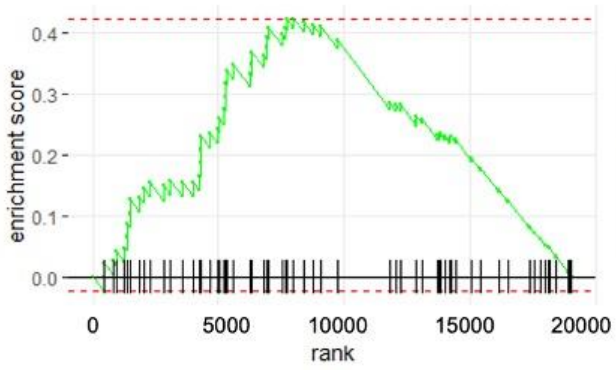
Figure 1. Heatmap of pathways in grasp group. The x axis is test pathways, and the y axis is SNPs. Numbers from 0 to 1 are indicated by color blue to white to red.

Pathway	Number of SNPs	Rate
hsa04976 Bile secretion	345	100%
hsa04150 mTOR signaling pathway	345	100%
hsa04722 Neurotrophin signaling pathway	345	100%
hsa00620 Pyruvate metabolism	345	100%
hsa04972 Pancreatic secretion	341	98.84%
hsa00785 Lipoic acid metabolism	333	96.52%
hsa04142 Lysosome	329	95.36%
hsa00770 Pantothenate and CoA biosynthesis	245	71.01%
hsa00563 Glycosyl phosphatidyl inositol (GPI)-anchor biosynthesis	136	39.42%
hsa00290 Valine, leucine and isoleucine biosynthesis	116	33.62%
hsa03010 Ribosome	92	26.67%
hsa05142 Chagas disease (American trypanosomiasis)	64	18.56%
hsa04114 Oocyte meiosis	26	7.54%
hsa05216 Thyroid cancer	20	5.80%
hsa04970 Salivary secretion	15	4.35%
hsa04141 Protein processing in endoplasmic reticulum	15	4.35%
hsa05214 Glioma	10	2.90%
hsa04146 Peroxisome	4	1.16%

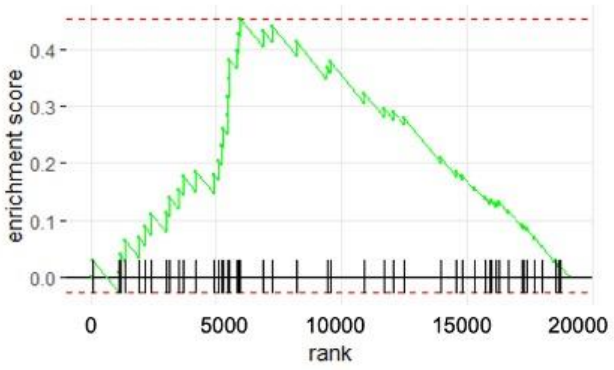
Table 1. Number of grasp SNPs that pathways show significance in. The first column is the name of the pathways and the pathway dataset used is "kegg.sets.hs" from "gageData" package. The second column is the number of SNPs that pathways show significance in. The total number of SNPs in grasp group is 345. The third column is the rate of pathways showing significance, which is calculated by number of SNPs divided by total number of SNPs (345).

Figure 2 shows the enrichment plots for the top eight pathways in Table 1, which are hsa04976 bile secretion, hsa04150 mTOR signaling pathway, hsa04722 neurotrophin signaling pathway, hsa00620 pyruvate metabolism, hsa04972 pancreatic secretion, hsa00785 lipoic acid metabolism, and hsa04142 lysosome. The green line shows how the enrichment score running for the gene set as the analysis goes along the ranked list.

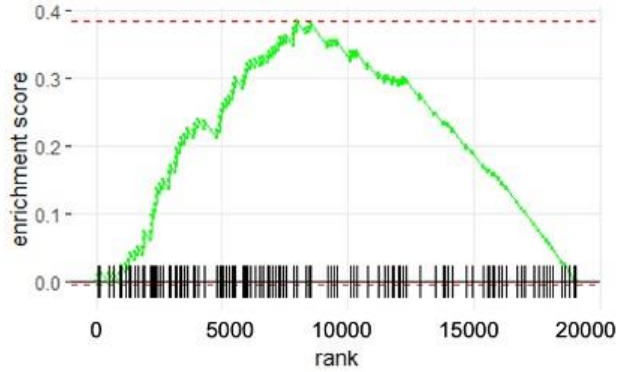
Figure 3 shows the enrichment plots for four pathways that have relatively smaller number of significant SNPs, which are hsa03010 ribosome, hsa05216 thyroid cancer, hsa04970 salivary secretion, and hsa04146 peroxisome. Pathway ribosome has 92 (26.67%) of significant SNPs. Pathway thyroid cancer has 20 (5.80%) of significant SNPs. Pathway salivary secretion has 15 (4.35%) of significant SNPs. Pathway peroxisome has 4 (1.16%) of significant SNPs.



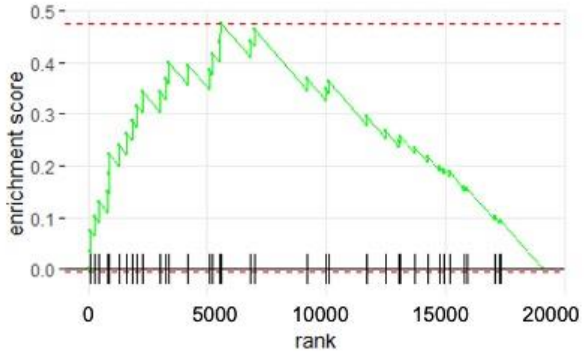
(a) hsa04976 Bile secretion



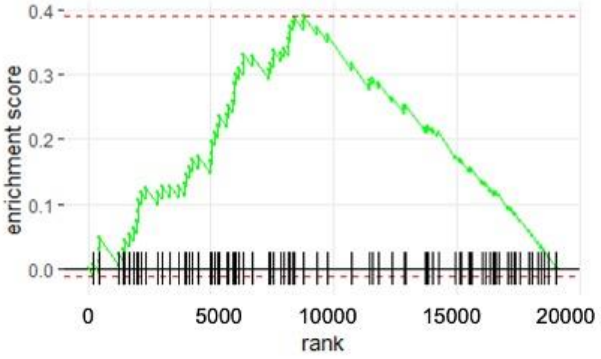
(b) hsa04150 mTOR signaling pathway



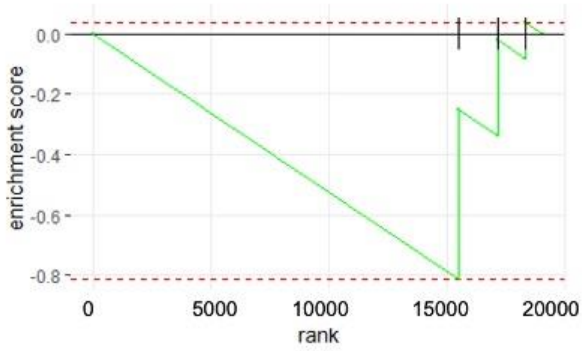
(c) hsa04722 Neurotrophin signaling pathway



(d) hsa00620 Pyruvate metabolism



(e) hsa04972 Pancreatic secretion



(f) hsa00785 Lipoic acid metabolism

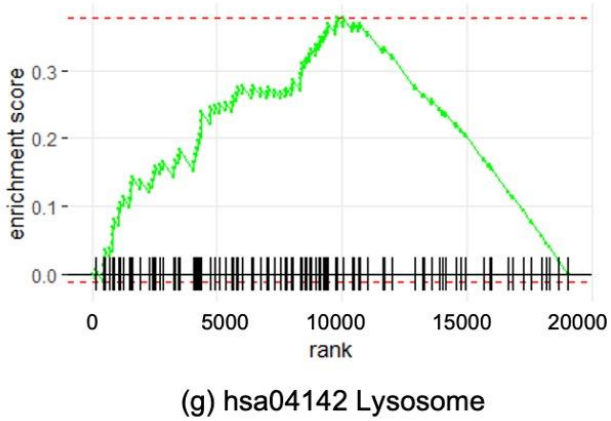


Figure 2. Enrichment Plots of Eight Pathways with High Rate of Significance. From (a) to (g), the pathways are hsa04976 bile secretion, hsa04150 mTOR signaling pathway, hsa04722 neurotrophin signaling pathway, hsa00620 pyruvate metabolism, hsa04972 pancreatic secretion, hsa00785 lipoic acid metabolism, and hsa04142 lysosome. The green line is the enrichment score of the rank. The red dotted line is the range of the enrichment score.

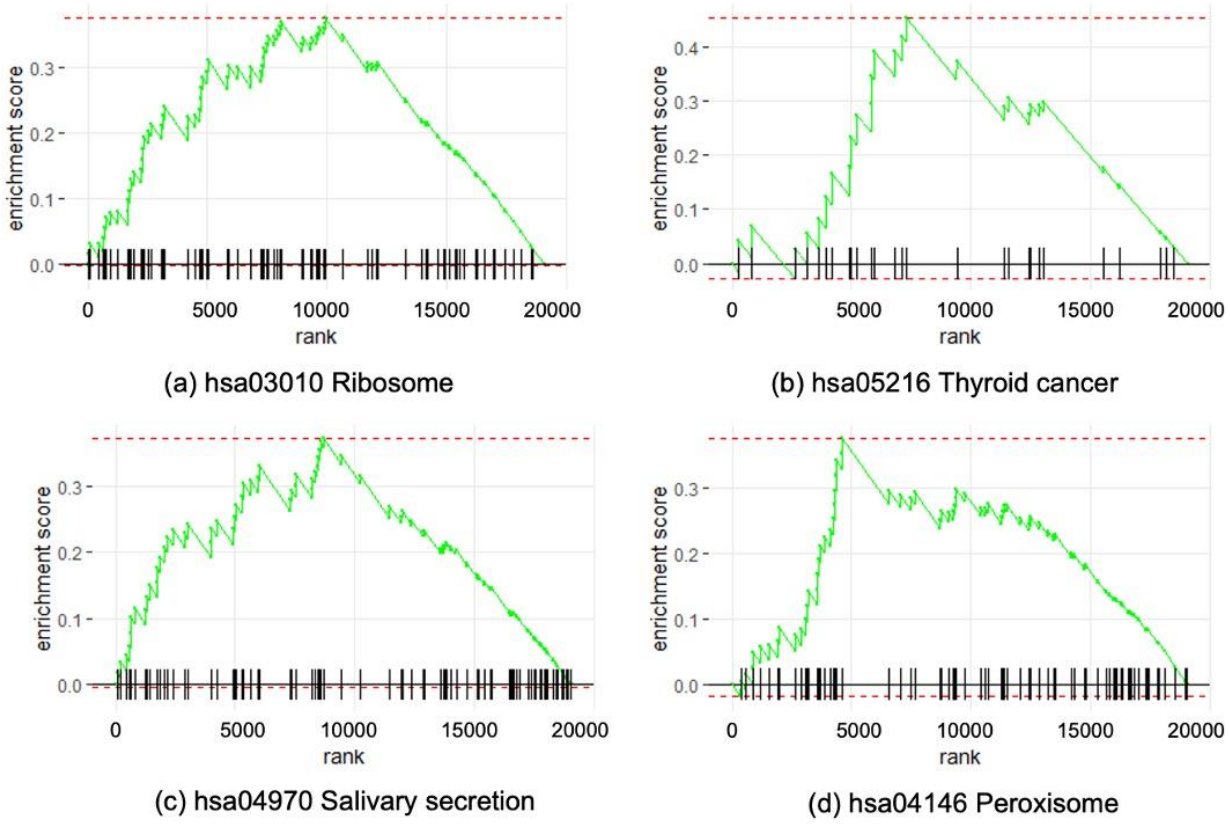


Figure 3. Enrichment Plots of Four Pathways with Small Number of Significant SNPs. From (a) to (d), hsa03010 ribosome, hsa05216 thyroid cancer, hsa04970 salivary secretion, and hsa04146 peroxisome. The green line is the enrichment score of the rank. The red dotted line is the range of the enrichment score.

Discussion

In the heatmaps, it can be found that some pathways have relatively consistent color while other pathways have different shade of color. The color indicates the p-values. The smaller the p-value is, the more it is close to dark blue. And oppositely, the bigger the p-value is, the more it is close to dark red. Some pathways have similar p-values across SNPs so that they have consistent color. The 18 pathways in Table 1 have difference extent of significance and the 7 pathways shown in the enrichment plots of Figure 2 have high rate of showing significance across SNPs, from 100% to 95.36%. In this case, we could say that the 7 pathways are related to Alzheimer's diseases. In Figure 2, the green line shows the enrichment score trend for the gene set as the analysis goes. The peak

points which are the furthest from 0 are the enrichment scores for the current gene set. The leading-edge subsets are the part from 0 to the rank at the max enrichment score, which contribute the most to the enrichment score [24]. All the pathways except lipoic acid metabolism in Figure 2(f) have enrichment score above 0, which means the 6 pathways are over expressed. However, the Figure 2(f) have negative enrichment score, which means the pathway of lipoic acid metabolism is under expressed. As we can see, Figure 2(f) has very few points of enrichment score, since the pathway lipoic acid metabolism only contains three genes: LIAS, LIPT1, and LIPT2. Small number of genes results in few points and drastic changes.

The 7 pathways in Figure 2 that show significance are: bile secretion, mTOR signaling pathway, neurotrophin signaling pathway, pyruvate metabolism, pancreatic secretion, lipoic acid metabolism, and lysosome. Studies have shown that genes involved in alternative bile acid synthesis pathways are expressed in the brain, while genes in the classical pathway are not [25]. Increased level of bile secretion is associated with Alzheimer's and cognitive decline. It's been known that the accumulation of Abeta and tau directly contribute to the cause of neurodegeneration, and mTOR is a target protein which is linked to the generation of Abeta and tau [26]. The over expression of mTOR signaling pathway is caused by PI3K/AKT pathway activated by Abeta administration [26]. Therefore, the up regulation of mTOR is a possible cause with Alzheimer's. Abeta and tau is not only associated with Alzheimer's, but also type-2 diabetes [27]. The impact of type-2 diabetes causes the increases of Abeta and amylin accumulation [27]. The up regulation of pancreatic secretion is linked with Alzheimer's. About neurotrophin family, there are two proteins being affected in the early phase of Alzheimer's: NGF and BDNF [28]. The over expression of neurotrophin signaling pathway increases the amount of NGF and BDNF, therefore related to the Alzheimer's. Among the Alzheimer's patients, the cerebral metabolic rate is impaired to some extent even before they show

any evidence of the clinical disease [29]. The impairment of pyruvate metabolism influences the glucose oxidation and, hence damaging the brain [29]. The dysfunction of lysosome is common among Alzheimer's patients which causes abnormal autophagy in cells. The abnormal process of autophagy results in the accumulation of autophagic vacuoles which should be resolved by lysosome [30]. The increase of lysosome may be one of the evidences of Alzheimer's and neurodegeneration. Lipoic acid is fatty acid that helps regulate metabolism [31]. The lack of lipoic acid can lead to dementia and neurodegeneration. Lipoic acid could be a potential treatment for Alzheimer's patients, but there is need to be aware of the possible harm if lipoic acid is provided in long term [31].

The four pathways in Figure 3 that show some significance are: ribosome, thyroid cancer, salivary secretion, and peroxisome. Protein dysfunction is found in the affected cortical regions of brains of patients with early-onset of Alzheimer's, and this is associated with ribosome dysfunction. The impairment of ribosome decreases the rate of protein synthesis. This may cause the over expression of ribosome to continue synthesize proteins needed for brain activities [32]. Thyroid cancer is related with the lack of iodine element. Researchs have found that iodine deficiencies can not only cause thyroid cancer but may also delay neuronal maturation and decrease brain weight [33]. The increasing expression may indicate that thyroid cancer can occur along with Alzheimer's. It's been known that tau protein and the phosphorylation of tau is associated with Alzheimer's. Saliva samples of Alzheimer's patients contain higher rate of phosphorylated tau compared to healthy control [34]. The over expression of salivary secretion may be an indicator of easier diagnosis of the Alzheimer's. Peroxisome proliferator-activated receptor gamma is involved in the regulation of inflammatory processes in the brain [35]. Considering the fact that inflammatory is related to Alzheimer's, the up-regulation of peroxisome may also be linked to Alzheimer's.

In conclusion, the 7 pathways with high rate of significance are associated with Alzheimer's disease. The 7 pathways are: bile secretion, mTOR signaling pathway, neurotrophin signaling pathway, pyruvate metabolism, pancreatic secretion, lysosome, and lipoic acid metabolism. The first 6 pathways are over expressed, and the last pathway lipoic acid metabolism is under expressed. In addition, the 4 pathways with smaller number of significance SNPs also give us some insights. The 4 pathways are: ribosome, thyroid cancer, salivary secretion, and peroxisome. These over expressed pathways may be regarded as potential causes or indicator of Alzheimer's.

Reference

[1]. Hachinski, V., & Munoz, D. G. (1997). Cerebrovascular pathology in Alzheimer's disease: cause, effect or epiphenomenon? *Annals of the New York Academy of Sciences*, 826(1), 1-6.

[2]. De la Torre, J. (1999). Critical threshold cerebral hypoperfusion causes Alzheimer's disease?. *Acta Neuropathol,*98, 1–8. https://doi.org/10.1007/s004010051044

[3]. The Centers for Disease Control and Prevention (CDC, 2019). Retrieved from https://www.cdc.gov/dotw/alzheimers/index.html.

[4]. Cummings, J. L., Cherry, D., Kohatsu, N. D., Kemp, B., Hewett, L., & Mittman B. (2002). Guidelines for Managing Alzheimer's Disease: Part II. Treatment. *Am Fam Physician.* 2002 Jun 15;65(12):2525-2534.

[5]. Basun, H., Bogdanovic, N., Ingelsson, M., et al. (2008). Clinical and Neuropathological Features of the Arctic APP Gene Mutation Causing Early-Onset Alzheimer Disease. *Arch Neurol.* 2008;65(4):499–505. https://doi.org/10.1001/archneur.65.4.499

[6]. Chung, J., Zhang, X., Allen, M. et al. Genome-wide pleiotropy analysis of neuropathological traits related to Alzheimer's disease. *Alz Res Therapy*. 10, 22 (2018). https://doi.org/10.1186/s13195-018-0349-z

[7]. Paaby, A. B., & Rockman, M. V. (2013). The many faces of pleiotropy. *Trends in genetics: TIG*, 29(2), 66–73. https://doi.org/10.1016/j.tig.2012.10.010

[8]. Ibanez L, Dube U, Davis AA, Fernandez MV, Budde J, Cooper B, Diez-Fairen M, Ortega-Cubero S, Pastor P, Perlmutter JS, Cruchaga C and Benitez BA (2018) Pleiotropic Effects of Variants in Dementia Genes in Parkinson Disease. *Front. Neurosci.* 12:230. https://doi.org/10.3389/fnins.2018.00230

[9]. Strittmatter, W. J., Weisgraber, K. H., Huang, D. Y., Dong, L. M., Salvesen, G. S., Pericak-Vance, M., et al. (1993). Binding of human apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. *Proceedings of the National Academy of Sciences*. Sep 1993, 90 (17) 8098-8102. https://doi.org/10.1073/pnas.90.17.8098

[10]. Karch, C. M. & Goate, A. M. (2015). Alzheimer's Disease Risk Genes and Mechanisms of Disease Pathogenesis. *Biological Psychiatry*: 77(1), 43-51. https://doi.org/10.1016/j.biopsych.2014.05.006

[11]. Ramanan, V. K., & Saykin, A. J. (2013). Pathways to neurodegeneration: mechanistic insights from GWAS in Alzheimer's disease, Parkinson's disease, and related disorders. *American journal of neurodegenerative disease*, 2(3), 145–175.

[12]. Cruchaga, C., Kauwe, J. S. K., Harari, O., Jin, S. H., Cai, Y., Karch, C. M., et al. (2013). GWAS of Cerebrospinal Fluid Tau Levels Identifies Risk Variants for Alzheimer's Disease. *Neuron*: 78(2), 256-268. https://doi.org /10.1016/j.neuron.2013.02.026

[13]. Cruchaga C, Chakraverty S, Mayo K, Vallania FLM, Mitra RD, et al. (2012) Rare Variants in APP, PSEN1 and PSEN2 Increase Risk for AD in Late-Onset Alzheimer's Disease Families. *PLoS ONE*, 7(2): e31039. https://doi.org/10.1371/journal.pone.0031039

[14]. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences,* 102(43), 15545-15550. https://doi.org/10.1073/pnas.0506580102

[15]. ROSMAP GWAS | NIAGADS. (2019). https://www.niagads.org/datasets/ng00029

[16]. De Jager, P. L., Shulman, J. M., Chibnik, L. B., Keenan, B. T., Raj, T., Wilson, R. S., et al. (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive

decline. *Neurobiology of Aging*, 33(5), 1017.e1-1017.e15. https://doi.org/10.1016/j.neurobiolaging.2011.09.033
[17]. Ensembl.org. (2020).
[18]. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC (2007)
[19]. PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.
[20]. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., Smyth, G. K. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, 43(7), e47. https://doi.org/10.1093/nar/gkv007
[21]. Sergushichev A (2016). "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation." bioRxiv. Doi:10.1101/060012, http://biorxiv.org/content/early/2016/06/20/060012.
[22]. Luo W (2019). gageData: Auxillary data for gage package. R package version 2.24.0.
[23]. Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., et al. (2020). gplots: Various R Programming Tools for Plotting Data. R package version 3.0.3.
[24]. Gene Set Enrichment Analysis (GSEA) User Guide. GSEA-msigdb.org. (2019).
[25]. Nho, K., Kueider-Paisley, A., MahmoudianDehkordi, S., Arnold, M., Risacher, S. L., Louie, G., et al. (2019). Altered bile acid profile in mild cognitive impairment and Alzheimer's disease: Relationship to neuroimaging and CSF biomarkers. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 15(2), 232–244. https://doi.org/10.1016/j.jalz.2018.08.012
[26]. Oddo S. (2012). The role of mTOR signaling in Alzheimer disease. *Frontiers in bioscience (Scholar edition)*, 4, 941–952. https://doi.org/10.2741/s310
[27]. Miklossy, J., & McGeer, P. L. (2016). Common mechanisms involved in Alzheimer's disease and type 2 diabetes: a key role of chronic bacterial infection and inflammation. *Aging*, 8(4), 575–588. https://doi.org/10.18632/aging.100921
[28]. Allen, S. J., Watson, J. J., & Dawbarn, D. (2011). The neurotrophins and their role in Alzheimer's disease. *Current neuropharmacology*, 9(4), 559–573. https://doi.org/10.2174/157015911798376190
[29]. Gray, L.R., Tompkins, S.C. & Taylor, E.B. Regulation of pyruvate metabolism and human disease. *Cell. Mol. Life Sci.* 71, 2577–2604 (2014). https://doi.org/10.1007/s00018-013-1539-2
[30]. McBrayer, M., & Nixon, R. A. (2013). Lysosome and calcium dysregulation in Alzheimer's disease: partners in crime. *Biochemical Society transactions*, 41(6), 1495–1502. https://doi.org/10.1042/BST20130201
[31]. Dos Santos, S. M., Romeiro, C., Rodrigues, C. A., Cerqueira, A., & Monteiro, M. C. (2019). Mitochondrial Dysfunction and Alpha-Lipoic Acid: Beneficial or Harmful in Alzheimer's Disease?. *Oxidative medicine and cellular longevity*, 2019, 8409329. https://doi.org/10.1155/2019/8409329
[32]. Ding, Q., Markesbery, W. R., Chen, Q., Li, F., & Keller, J. N. (2005). Ribosome dysfunction is an early event in Alzheimer's disease. The Journal of neuroscience: the official journal of the Society for Neuroscience, 25(40), 9171–9175. https://doi.org/10.1523/JNEUROSCI.3040-05.2005
[33]. Foster, H. D. (1987). Disease family trees: The possible roles of iodine in goitre, cretinism, multiple sclerosis, amyotrophic lateral sclerosis, Alzheimer's and Parkinson's

diseases and cancers of the thyroid, nervous system and skin. *Medical Hypotheses*, 24(3), 249-263. https://doi.org/10.1016/0306-9877(87)90072-7

[34]. Shi, M., et al. (2011). Salivary Tau Species Are Potential Biomarkers of Alzheimer's Disease. Journal of Alzheimer's Disease, 27(2), 299-305. https://doi.org/ 10.3233/JAD-2011-110731

[35]. Scacchni, R., Pinto, A., Gambina, G., Rosano, A. & Corbo, R. M. (2007). The peroxisome proliferator-activated receptor gamma (PPAR-γ2) Pro12Ala polymorphism is associated with higher risk for Alzheimer's disease in octogenarians. Brain Research, 1139, 1-5. https://doi.org/10.1016/j.brainres.2006.12.078