Effects of Genetic Variation on the Dynamics of Neurodegeneration in Alzheimer's Disease

Blake P. Printy, Nishant Verma, Matthew C. Cowperthwaite, Mia K. Markey, *Senior Member, IEEE*, for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*— Although many genetic markers are identified as being associated with Alzheimer's disease (AD), not much is known about their association with the structural changes that happen as the disease progresses. In this study, we investigate the genetic etiology of neurodegeneration in AD by associating genetic markers with atrophy profiles obtained using patient data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The atrophy profiles were quantified using a linear least-squares regression model over the span of patient enrollment, and used as imaging features throughout the analysis. A subset of the imaging features were selected for genetic association based on their ability to discriminate between healthy individuals and AD patients in a Support Vector Machines (SVM) classifier. Each imaging feature was associated with single-nucleotide polymorphisms (SNPs) using a linear model that included age and cognitive impairment scores as covariates to correct for normal disease progression. After false discovery rate correction, we observed 53 significant associations between SNPs and our imaging features, including associations of ventricular enlargement with SNPs on estrogen receptor 1 (ESR1) and sortilin-related VPS10 domain containing receptor 1 (SORCS1), hippocampal atrophy with SNPs on ESR1, and cerebral atrophy with SNPs on transferrin (TF) and amyloid beta precursor protein (APP). This study provides important insights into genetic predictors of specific types of neurodegeneration that could potentially be used to improve the efficacy of treatment strategies for the disease and allow the development of personalized treatment plans based on each patient's unique genetic profile.**

I. INTRODUCTION

Alzheimer's disease (AD), the most common primary neurodegenerative dementia, is a genetically complex disorder that affects an estimated 5.3 million people in the United States [1]. As the size of the world's elderly population increases, AD will become an even more devastating public health problem and a more significant economic burden. Characterized by progressive cognitive impairment, AD results from continuous neurodegeneration in specific regions of the brain. Structural changes such as hippocampal atrophy, ventricular enlargement, and cortical atrophy have been shown to support the diagnosis of AD, and recently have been used in diagnostic models for discriminating between AD patients and healthy controls [2].

Since AD is caused by a complex set of genetic and neurophysiological factors, it has been difficult for researchers to understand the underlying causes of neuronal loss for specific brain regions. The molecular basis of AD is characterized by the formation of two main protein aggregates – senile plaques and neurofibrillary tangles – which are involved in progressive neuronal degeneration and death. Senile plaques formed in AD are generated by a deposition of fibrils of the Aβ peptide, a fragment derived from proteolytic processing of the amyloid beta precursor protein (APP) [3]. Several genes are involved in regulating APP processing and thus have much potential to explain the biochemical nature of neuronal loss observed throughout disease progression. Many genes have been shown to be associated with AD in addition to the well-known risk factor APOE-ε4, but were typically identified on the basis of a binary association between the disease and intergenic or nearby SNPs.

Although the structural dynamics of AD are extensively studied, not much work has been done to relate structural changes in the brain to genetic variants. Previous efforts to associate magnetic resonance imaging-based imaging biomarkers with genetic markers have generally focused on a small number of genetic variables (often only the presence of the APOE$-\varepsilon$4 allele) [4], whereas genome-wide association studies (GWAS) of AD have typically only examined a small set of imaging phenotypes [5]. Unfortunately, these studies are only able to provide focused insights into the nature of structural changes in AD because they do not take full advantage of the genetic and imaging information available in modern datasets.

Considering that substantial brain damage has already occurred by the time AD is typically diagnosed, it is crucial to develop diagnostic strategies that can predict progression to AD (or a similar type of dementia) at an early, prodromal stage of mild cognitive impairment. To this end, we investigated the potential of using genetic risk factors in developing genetically-derived models of neurodegeneration in AD. Such models would not only improve understanding of the biological nature of the disease, but would offer great perspective to researchers developing treatment strategies for AD. Moreover, such models would also be helpful in improving the efficacy of clinical trials investigating disease-

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database www.loni.ucla.edu/ADNI. As such, the investigators within the ADNI did not participate in the analysis or writing of this report.

B. P. Printy is with Dept. of Biomedical Engineering, The University of Texas at Austin, TX 78712, USA, blakeprinty@utexas.edu

N. Verma is jointly affiliated with Dept. of Biomedical Engineering, The University of Texas at Austin and NeuroTexas Institute, St. Davids HealthCare, Austin, vnishant@utexas.edu

M. C. Cowperthwaite is with the Texas Advanced Computing Center, The University of Texas at Austin, TX 78758, USA, Phone: 512-475-9411, mattcowp@tacc.utexas.edu

M. K. Markey is with Dept. of Biomedical Engineering, The University of Texas at Austin, TX 78712, USA, and Dept. of Imaging Physics, The University of Texas MD Anderson Cancer Center, Phone: 512-471-1711, mia.markey@utexas.edu

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

modifying drugs for AD.

II. DATA

A. Imaging Data

The data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [6]. T1-weighted MRI scans were provided for 217 healthy controls (NL), 361 patients with moderate to severe mild cognitive impairment (MCI), and 179 patients with moderate to severe AD at 0-, 6-, 12-, 18-, 24- and 36-month time intervals. Each of the images provided by the ADNI had previously been preprocessed using steps to correct for image geometry distortion and non-uniform voxel intensity distributions. For more information on the imaging protocols enforced by ADNI, see ADNI's homepage `adni.loni.ucla.edu`.

B. Genetic Data

All ADNI participants were genotyped with the Illumina Human610-Quad BeadChip (Illumina, Inc., San Diego, CA) platform, which determined the genotypes of 620,901 SNP and CNV markers. Data processing and genotype calling was performed by ADNI sites using BeadStudio, Illumina GenomeStudio v2009.1. The two SNPs (rs429358, rs7412) that define the $\varepsilon 2$, $\varepsilon 3$, and $\varepsilon 4$ alleles of the apolipoprotein E gene (APOE) (the strongest known genetic risk factor for AD), are not on the Human610-Quad BeadChip, and thus were not considered in this study. All genotype data was obtained from ADNI in ped format, which allowed for data management and processing in the open-source tool PLINK `http://pngu.mgh.harvard.edu/~purcell/plink/` PLINK [7].

C. Patient Assessments

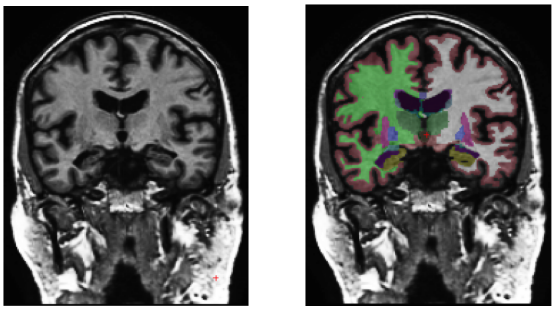
Along with the imaging and genetic data, clinical assessments of each patient's cognitive impairment were also obtained from ADNI. During each visit, the Alzheimer's Disease Assessment Scale (ADAS-cog) and Mini-Mental State Examination (MMSE) scores were collected to measure disease progression throughout the study.

III. METHODS

A. Image Segmentation

Volumetric segmentation was performed using the FreeSurfer image analysis suite `http://surfer.nmr.mgh.harvard.edu/`. Briefly, the FreeSurfer pipeline includes the following: correction for motion artifacts by averaging multiple volumetric T1 weighted images, a hybrid watershed/surface deformation procedure to remove non-brain tissue, an affine transformation into Talairach space, an intensity normalization procedure, segmentation of volumetric subcortical deep gray matter and white matter, tessellation of gray and white matter structural boundaries, and application of intensity gradients to optimally place gray/cerebrospinal and gray/white fluid boundaries at locations where large shifts in intensity define transitions across tissue classes [8]. Procedures for measuring cortical thickness have been validated against manual measurements [9]. FreeSurfer morphometric procedures have been demonstrated to show good test-retest reliability across scanner manufacturers and across field strengths [10].

To extract reliable volume and thickness estimates, images were automatically processed with the longitudinal stream [10] in FreeSurfer. Explicitly, an unbiased within-subject image and template space is constructed using robust, inverse registration [10]. Several processing steps, including Talairach transformation, atlas registration, skull stripping, and spherical surface mapping and parcellation are then initialized with common information from each within-subject template, which work to significantly increase reliability and statistical power [10].



(a) Original MRI volume (b) Labelled MRI volume

Fig. 1. Sample MRI image slice and corresponding Freesurfer labeling after longitudinal pipeline.

B. Feature Quantification

After labeling images with the longitudinal pipeline and quantifying volumes using FreeSurfer, atrophy profiles were quantified for each segmented volume using a linear least-squares regression model over the span of patient enrollment. Slopes obtained from the model were used as imaging features throughout the analysis, representing rates of neurodegeneration over time.

C. Feature Selection

To minimize the dimensionality of the imaging feature set into a subset of highly informative quantitative traits for association, a classification model was built to discriminate between healthy individuals and AD patients using subsets of imaging features. The model was iteratively tested with increasingly large subsets of features that were selected using the commonly employed maximum-relevance minimum-redundancy (mRMR) algorithm [11]. The mRMR algorithm uses mutual information between two groups x and y, $I(x, y)$, as a measure of closeness, and tries to iteratively choose features that are maximally informative (i.e. being able to discriminate between diagnostic groups) and at the same time minimally redundant with respect to the remainder of features in the candidate set. Mutual information for the continuous imaging features was quantified using kernel density estimation with a bandwidth derived from mean integrated squared error (MISE). The relevance, D,

and redundancy, R, criteria are respectively quantified with respect to a patient's binary diagnostic status ($NL = 0$, $AD = 1$), c, and the features, x_i, in the feature space, S as follows:

$$D(S,c) = \frac{1}{|S|} \sum_{x_i \epsilon S} I(x_i, c)$$

$$R(S) = \frac{1}{|S|^2} \sum_{x_i x_j \epsilon S} I(x_i, x_j)$$

Feature optimality was quantified in the algorithm as the difference between relevance and redundancy.

$$\Phi = D - R$$

Features were iteratively removed from the candidate set until a desired number of features was reached. The following illustrates implementation of the mRMR algorithm for each iteration ($X - S_{m-1}$ is the updated feature set after each iteration m).

$$\max_{x_j \epsilon X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \epsilon S_{m-1}} I(x_j, x_i) \right]$$

To assess the discriminatory power of selected feature subsets, a nonlinear support vector machine (SVM) was developed and tested under 5-fold cross-validation using increasing numbers of features. During each round of testing, stratified sampling was used to split the data into feature selection (2/3 of the data) and training/testing sets (1/3 of the data). The mRMR algorithm was then used to select a desired number of informative features, and the selected features were used to train and test an SVM classifier that was developed using the C-SVC formulation and a Laplacian kernel. A Laplacian kernel was chosen because of its performance in identifying non-linear relationships. Mean area under the Receiver Operating Characteristic curve (AUC) was used to quantify performance of the feature subsets in discriminating between healthy controls and AD patients, and an optimal feature subset was selected for genetic association based on converging performance.

D. SNP Selection

To minimize the effects of multiple comparisons in the association analysis, an optimally relevant set of SNPs was selected. A set of 66 candidate genes was selected based on literature searches, the Online Mendelian Inheritance in Man (OMIM) database, and the AlzGene database [12,13]. SNPs within the physical regions occupied by the candidate genes were extracted from data downloaded from Ensembl's Biomart utility [14].

Several inclusion thresholds for the data were implemented using PLINK [7]. Individuals with more than 10% missing genotypes were removed from the analysis, along with SNPs having a minor allele frequency below 5% or a call rate below 90%. SNPs not in Hardy-Weinberg Equilibrium ($p<0.001$) were also excluded from the analysis. These inclusion thresholds yielded a set of 599 SNPs for the association analysis.

E. SNP/Phenotype Associations

In order to identify SNPs associated with rates of neurodegeneration, two groups were used in the analysis: patients showing some degree of cognitive impairment throughout their enrollment, and healthy controls.

A multiple linear regression model was used to determine if SNP allele frequency had any significant effects on atrophy profiles. Letting X_i represent the allele frequency of the examined SNP for patient i, V_i represent the set of all visits for patient i, Y_{i,t_j} represent the ADAS-cog score for patient i at time point t_j, and Z_{i,t_j} represent the age for patient i at time point t, the model is defined as:

$$y = \beta_0 + \beta_1 X_i + \beta_2 \left(\frac{1}{|V_i|} \sum_{t_j \epsilon V_i} Y_{i,t_j} \right) + \beta_3 \left(\frac{1}{|V_i|} \sum_{t_j \epsilon V_i} Z_{i,t_j} \right)$$

The covariates in the model are the mean ADAS-cog score (β_2) and mean age (β_3) of each patient throughout their enrollment in the study. Gender was also tested as a covariate, but did not add any power to the model (not shown). All association testing was performed with PLINK.

IV. RESULTS

A. Classifier Performance with Selected Features

Fig. 1 details performance of the classification model developed to select features for the association analysis. Classification performance for the model converged at approximately 12 features, and these top 12 features were used as quantitative phenotypic traits in the genetic-association analysis.

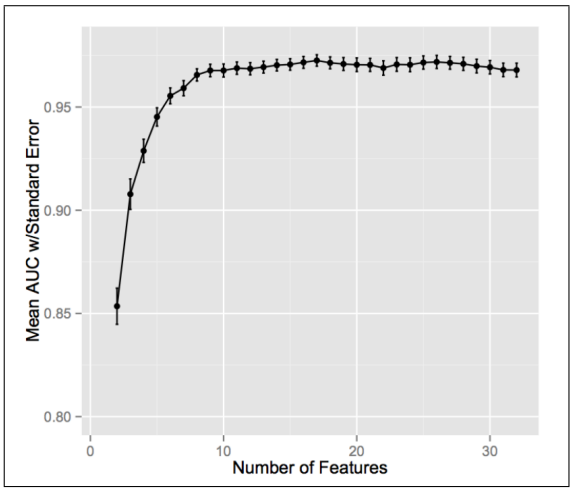


Fig. 2. Performance of the SVM classifier for different subsets of features selected by the mRMR algorithm.

B. Significant Associations

Table I shows the top 10% of the most significant associations, where all p-values shown are corrected using the false discovery rate (FDR) control. Table II shows a summary of significant associations ($p < 0.05$) by gene and imaging feature, where emphasis is placed on the number of unique SNPs identified for each gene.

TABLE I

TOP 10% OF SIGNIFICANT ASSOCIATIONS BY SNP IDENTIFIER. ALL P-VALUES SHOWN ARE CORRECTED USING THE FALSE DISCOVERY RATE (FDR) CONTROL.

Variation	Gene	Gene Description	Imaging Feature	P-Value
rs9341052	ESR1	Estrogen receptor 1	Left Lateral Ventricular Enlargement	2.513×10^{-5}
rs4726618	EPHA1	EPH receptor A1	Left Inferior Lateral Ventricular Enlargement	2.985×10^{-4}
rs9341052	ESR1	Estrogen receptor 1	Right Lateral Ventricular Enlargement	1.186×10^{-3}
rs17014923	BIN1	Bridging integrator 1	Third Ventricular Enlargement	2.449×10^{-3}
rs6584777	SORCS1	Sortilin-related VPS10 domain containing receptor 1	Left Inferior Lateral Ventricular Enlargement	4.407×10^{-4}
rs749008	BIN1	Bridging integrator 1	Third Ventricular Enlargement	5.694×10^{-3}

TABLE II

SIGNIFICANT ASSOCIATIONS BY GENE AND UNIQUE SNP COUNT.

Imaging Feature	Gene	Unique SNPs
Lateral Ventricular Enlargement	ESR1	6
	BIN1	1
	LDLR	1
Inferior Lateral Ventricular Enlargement	SORCS1	10
	ESR1	3
	APP	2
Hippocampal Atrophy	ESR1	18
	LRAT	3
	APP	1
Cortical Atrophy	TF	4
	APP	2
	SORCS1	2

V. DISCUSSION

We identified several SNP loci that are associated with specific types of neurodegeneration in specific brain regions of AD patients. Several of the SNPs found to be associated with imaging features are in or adjacent to genes that have been previously shown to be associated with biochemical markers for AD [15,16]. In particular, genetic variation on SORSC1 has been shown to alter Aβ protein processing [15], and estrogen treatment affecting ESR1 has been shown to modulate the risk of developing AD in women [16].

The significant associations found in this study further stress the need for investigating genetic factors associated with structural changes in AD. These SNP associations suggest that genetic variation can potentially be useful to diagnostics for identifying particular types of neurodegeneration in AD, and thus highlight the need to include genetic variation in any diagnostic models for predicting the onset of AD symptoms.

VI. FUTURE WORK

Future plans for this study include the development of statistical models for quantifying disease severity and diagnosing AD at a prodromal stage of MCI using the investigated imaging and genetic features. Also, we plan to further investigate SNPs found significant in this study on a pathway-level basis, in order to understand more about the biological nature of neurodegeneration in AD.

ACKNOWLEDGMENT

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering.

REFERENCES

[1] Hebert et al., "Alzheimer disease in the United States (20102050) estimated using the 2010 census," Neurology, Feb. 2013.

[2] M. Benoit et al., "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," Neuroradiology, vol. 51, pp. 73-83, 2009.

[3] R. Swerdlow, "Pathogenesis of Alzheimer's disease," Clin Interv Aging, vol. 2, no. 3, pp. 347-359, Sep. 2007.

[4] S. G. Potkin et al., "Hippocampal atrophy as a quantitative trait in genome-wide association study identifying novel susceptibility genes for Alzheimer's disease," PLoS One, vol. 4, no. 8, pp. e6501, Aug. 209.

[5] F. Nicola et al., "Anatomicaly-distinct genetic associations of APOE $\varepsilon 4$ allele load with regional cortical atrophy in Alzheimer's disease," Neuroimage, vol. 44, no. 1, pp. 724-728, Feb. 2009.

[6] Mueller et al., "The Alzheimers Disease Neuroimaging Initiative," Neuroimaging Clin. N. Am., vol. 15, no. 4, pp. 869xii, Nov. 2005.

[7] S. Purcell et al., "PLINK: a toolset for whole-genome association and population-based linkage analysis," American Journal of Human Genetics, vol. 81, 2007.

[8] A. M. Dale, B. Fischl, M. I. Sereno, "Cortical surface-based analysis. I. Segmentation and surface reconstruction," Neuroimage, vol. 9, pp. 179-194, 1999.

[9] D. H. Salat et al., "Thinning of the cerebral cortex in aging," Cereb Cortex, vol. 14, pp. 721-730, 2004.

[10] M. Reuter et al., "Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis," Neuroimage, vol. 61, no. 4, pp. 1402-1418, 2012. [Online]. Available: http://reuter.mit.edu/papers/reuter-long12.pdf

[11] H. Peng, C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[12] McKusick-Nathans Institute of Genetic Medicine, "Online Mendelian Inheritance in Man, OMIM," Internet: http://omim.org/. html, Jan. 14, 2014 [Feb. 12, 2014].

[13] L. Bertram et al., "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database." Nat Genet, vol. 39, no. 1, pp. 17-23, 2007.

[14] P. Flicek et al., "Ensembl 2014," Nucleic Acids Research, vol. 42(Database Issue), 2014.

[15] C. Reitz et al., "SORCS1 alters amyloid precurser protein processing and variants may increase Alzheimer's disease risk," Ann Neurol, vol. 69, no. 1, pp. 47-67, Jan. 2011.

[16] L. Goumidi et al., "Study of estrogen receptor alpha and receptor beta gene polymorphisms on Alzheimer's disease," J Alzheimers Dis, vol. 26, no. 3, pp. 431-439, 2011.