bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Multivariate association between single-nucleotide polymorphisms in Alzgene linkage regions and structural changes in the brain: discovery, refinement and validation

ELENA SZEFER, DONGHUAN LU, FAROUK NATHOO, MIRZA FAISAL BEG, JINKO GRAHAM,*
for the Alzheimers Disease Neuroimaging Initiative†
Department of Statistics and Actuarial Science,
Simon Fraser University,
8888 University Dr, Burnaby, BC, Canada V5A 1S6
jgraham@sfu.ca

July 30, 2017

Abstract

Both genetic variants and brain region abnormalities are recognized to play a role in cognitive decline. We explore the association between single-nucleotide polymorphisms (SNPs) in linkage regions for Alzheimer's disease and rates of decline in brain structure using data from the Alzheimers Disease Neuroimaging Initiative (ADNI).

*To whom correspondence should be addressed.

†Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

In an initial discovery stage, we assessed the presence of linear association between the minor allele counts of 75,845 SNPs in the Alzgene linkage regions and predicted rates of change in structural MRI measurements for 56 brain regions using an RV test. In a second, refinement stage, we reduced the number of SNPs using a bootstrap-enhanced sparse canonical correlation analysis (SCCA) with a fixed tuning parameter. Each SNP was assigned an importance measure proportional to the number of times it was estimated to have a nonzero coefficient in repeated re-sampling from the ADNI-1 sample. We created refined lists of SNPs based on importance probabilities greater than 50% and 90%, respectively. In a third, validation stage, we assessed the multivariate association between these refined lists of SNPs and the rates of structural change in the independent ADNI-2 study dataset.

There was strong statistical evidence for linear association between the SNPs in the Alzgene linkage regions and the 56 imaging phenotypes in both the ADNI-1 and ADNI-2 samples ($p < 0.0001$). The bootstrap-enhanced SCCA identified 1,694 priority SNPs with importance probabilities $> 50\%$ and 22 SNPs with importance probabilities $> 90\%$. The 1,694 prioritized SNPs in the ADNI-1 data were associated with imaging phenotypes in the ADNI-2 data ($p = 0.0021$).

This manuscript presents an analysis that addresses challenges in current imaging genetics studies such as biased sampling designs and high-dimensional data with low-signal. Genes corresponding to priority SNPs having the highest contribution in the validation data have previously been implicated or hypothesized to be implicated in AD, including GCLC, IDE, and STAMBP1andFAS. We hypothesize that the effect sizes of the 1,694 SNPs in the priority set are likely small, but further investigation within this set may advance understanding of the missing heritability in late-onset Alzheimers disease. Multivariate analysis; Linkage regions; Imaging genetics; Endophenotypes; Inverse probability weighting; Variable importance probabilities

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder causing cognitive impairment and memory loss. The estimated heritability of late-onset AD is 60%-80% (Gatz et al., 2006), and the largest susceptibility allele is the $\varepsilon 4$ allele of *APOE* (Corder et al., 1993), which may play a role in 20% to 25% of AD cases. Numerous studies have identified susceptibility genes which account for some of the missing heritability of AD, with many associated variants having been identi-

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

fied through genome-wide association studies (GWAS) (e.g. Beecham et al., 2009; Kamboh et al., 2012; Bertram et al., 2008). Apart from *APOE*, the associated variants have mostly had moderate or small effect sizes, suggesting that the remaining heritability of AD may be explained by many additional genetic variants of small effect.

Identifying susceptibility variants with small effect sizes in GWAS is challenging since strict multiple testing corrections are required to maintain a reasonable family-wise error rate. This analysis focuses on leveraging information from prior family of studies of AD (Hamshere et al., 2007; Butler et al., 2009), by looking for association in previously identified linkage regions reported on the Alzgene website (Biomedical Research Forum, 2013). Linkage regions for AD are genomic regions that tend to be co-inherited with AD in families. By definition, linkage regions include susceptibility genes that are co-transmitted with the disease. The regions currently identified from family studies of AD are large, however, since families contain relatively few transmissions. Further transmissions over multiple generations would provide more fine-grain information about the location of susceptibility genes. Previous studies have fine-mapped a single linkage region through association of AD with genetic variants in densely genotyped or sequenced regions (Fallin et al., 2010; Ertekin-Taner, 2003; Scott et al., 2000; Züchner et al., 2008), or have confirmed linkage to AD in genomic regions identified from GWAS (Anna et al., 2011) . In this report, we aim to fine-map multiple linkage regions for AD through multivariate association of their SNPs to the rates of atrophy in brain regions affected by AD.

We analyze data from two phases of the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005), ADNI-1 and ADNI-2, which are case-control studies of AD and mild-cognitive impairment. The rates of atrophy in brain regions affected by AD are so-called endophenotypes: observable traits that reflect disease progression. By investigating the joint association between the genomic variants and the neuroimaging endophenotypes, we use the information about disease progression to supervise the selection of single-nucleotide polymorphisms (SNPs). This multivariate approach to analysis stands in contrast to the commonly-used mass-univariate approach (see Nathoo et al. (2017) for a review of the mass-univariate analysis in imaging genetics studies) in which separate regressions are fit for each SNP, and the disease outcome is predicted by the minor allele counts. Simultaneous analysis of association is preferred because the reduced residual variation leads to (i) a clearer assessment of the signal from each SNP, (ii) increased power to detect signal, and (iii) a decreased false-positive rate (Hoggart et al., 2008). We also employ inverse probability weighting to account

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for the biased sampling design of the ADNI-1 and ADNI-2 studies, an aspect of analysis that has not been accounted for in many previous imaging genetics studies (Zhu et al., 2016).

Methods that explicitly account for gene structure have been proposed for analyzing the association between multiple imaging phenotypes and SNPs in candidate genes (e.g. Wang et al., 2011; Greenlaw et al., 2017). However, these methods become computationally intractable when analyzing data with tens of thousands of genotyped variants. To select SNPs associated with disease progression, we instead use sparse canonical correlation analysis (SCCA) to find a sparse linear combination of SNPs having maximal correlation with the imaging endophenotypes. Multiple penalty schemes have been proposed to implement the sparse estimation in SCCA (Parkhomenko et al., 2009; Witten et al., 2009; Lykou and Whittaker, 2010). We employ an SCCA implementation that estimates the sparse linear combinations by computing sparse approximations to the left singular vectors of the cross-correlation matrix of the SNP data and the neuroimaging endophenotype data (Parkhomenko et al., 2009). Sparsity is introduced through soft-thresholding of the coefficient estimates (Donoho and Johnstone, 1994), which has been noted (Chalise and Fridley, 2012) to be similar in implementation to a limiting form of the elastic-net (Zou and Hastie, 2005). A drawback of ℓ_1-type penalties is that not all SNPs from an LD block of highly-correlated SNPs that are associated with the outcome will be selected into the model (Zou and Hastie, 2005). We prefer an elastic-net-like penalty over alternative implementations with ℓ_1 penalties because it allows selection of all potentially associated SNPs regardless of the linkage-disequilibrium (LD) structure in the data.

We may think of SNP genotypes as a matrix X and imaging phenotypes as a matrix Y measured on the same n subjects. Robert and Escoufier (1976) showed that estimating the maximum correlation between linear combinations of X and Y in canonical correlation analysis is equivalent to estimating the linear combinations having the maximum RV coefficient, a measure of linear association between the multivariate datasets (Escoufier, 1973). As the squared correlation coefficient between the first canonical variates, the RV coefficient is well-suited for testing linear association in our context. We use a permutation test based on the RV coefficient to assess the association between the initial list of SNPs in X and the phenotypes in Y. A permutation test with the RV coefficient is preferred over a parametric hypothesis test since the permutation null distribution is computed under the same conditions as the observed RV coefficient, resulting in a valid hypothesis test. The outcome of this test is used to determine whether or not to proceed with a second refinement stage that reduces the number of SNPs by ap-

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

plying SCCA.

Tuning parameter selection can be difficult when data has low signal (Nathoo et al., 2016) and selection of the soft-thresholding parameter in SCCA is challenging in our context. Since the number of SNPs exceeds the sample size and many of the SNPs are expected to be unassociated with the phenotypes, large sample correlations can arise by chance (Fan et al., 2011). Indeed, the prescribed procedure of selecting the penalty parameter with highest predicted correlation across cross-validation test sets (Parkhomenko et al., 2009) results in more than 98% of the SNPs remaining in the model. A prediction criterion for choosing the penalty term may contribute to the lack of variable selection, allowing redundant variables into the model (Leng et al., 2006). When the same tuning parameter is used for variable selection and shrinkage, redundant variables tend to be selected to compensate for overshrinkage of coefficient estimates and losses in predictive ability (Radchenko and James, 2008). In our case, there is effectively no variable selection and little insight is gained by allowing for sparsity in the solution. To circumvent the lack of variable selection from SCCA, we fix the tuning parameter to select about 10% of the SNPs (Wu et al., 2009) and then use resampling to determine the relative importance of each SNP to the association with neuroimaging endophenotypes. Instead of using the prediction-optimal penalty term, we fixed the soft-thresholding parameter for the SNPs to achieve variable selection based on the rationale that no more than about 7,500 SNPs, or approximately 10%, are expected to be associated with the phenotypes. This choice is guided by prior experience in genetic association studies, where the majority of genetic variants have no effect on the phenotypes, or an effect that is indistinguishable from zero (Carbonetto and Stephens, 2012).

The organization of the manuscript is as follows. The Materials and Methods section describes the ADNI data, the data processing procedures, and the methods applied for discovery, refinement, and validation. The Results section presents the results of the analyses. The Discussion section notes challenges and successes of the analysis, including considerations for modelling continuous phenotype data under a case-control sampling design, and provides interpretation of the results.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2 Materials and Methods

2.1 Materials

Figure 1 illustrates the data processing steps required to compute the quantities that are analyzed from the raw data: adjusted minor allele counts of SNPs in the Alzgene linkage regions for the genomic data, and adjusted predicted rates of change at 56 brain regions of interest (ROIs) from MRIs for the neuroimaging data. The following subsections detail the data processing steps required to obtain the analysis and validation datasets.

2.1.1 ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Patients with MCI have a subjective memory concern but no significant cognitive impairment, while patients with early Alzheimer's disease have experienced memory loss. The first phase of ADNI, ADNI-1, aimed to enroll approximately 800 study subjects, aged 55-90, of which 200 were cognitively normal (CN), 200 had mild Alzheimers disease (AD), and 400 had mild cognitive impairment (MCI) (ADNI Procedures Manual, 2006, p. 3). The CN subjects were roughly age-matched to the MCI and AD subjects. Subjects were assigned to a group following inclusion and exclusion criteria based on clinical and cognitive tests. Mild AD subjects had mini-mental state examination (MMSE) scores from 20-26 inclusive, clinical dementia rating (CDR) of 0.5 or 1.0, and met the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimers Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD. MCI subjects had MMSE scores between 24-30, CDR of 0.5, a memory complaint, and objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II (Wechsler, 2009). Cognitively normal participants were non-depressed, with MMSE scores between 24-30, and did not fit the criteria for the MCI or mild AD groups (ADNI Procedures Manual, 2006, p. 3-4). The study subjects were enrolled from 50 sites in the US and Canada and were required to commit to at least two years of follow up MRI.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

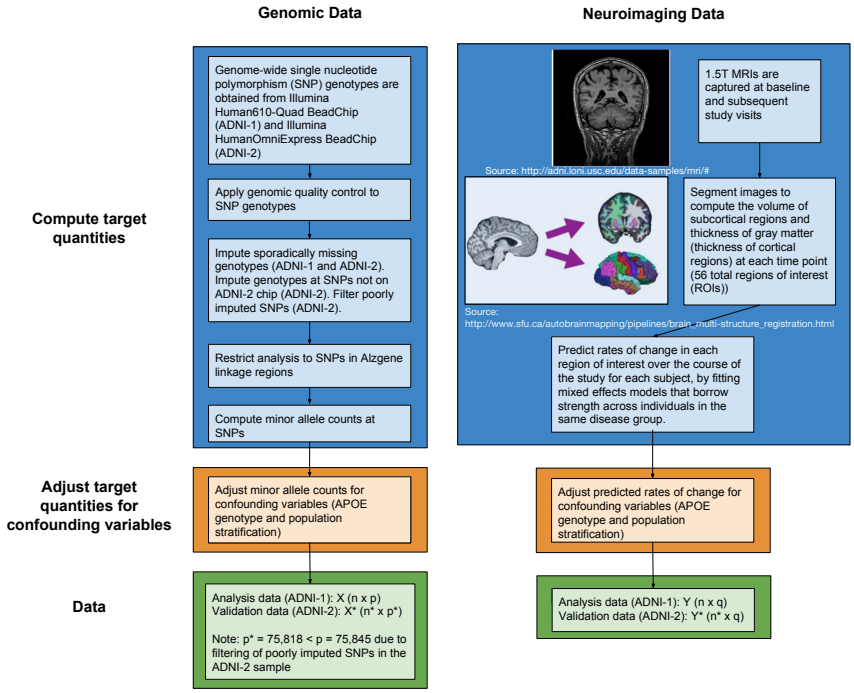

Figure 1: Flow chart of the steps to process the raw genomic and neuroimaging data into the analysis and validation datasets. The target quantities, minor allele counts at SNPs in Alzgene linkage regions for the genomic data and predicted rates of change at brain regions of interest for the neuroimaging data, are computed. Then, both sets of targets quantities are adjusted for potentially confounding variables to obtain the data for analysis (ADNI-1) or validation (ADNI-2).

ADNI-2 was a subsequent phase of ADNI that involved enrollment of new subject cohorts, as well as continued follow-up of subjects enrolled in earlier phases. ADNI-2 enrolled a new cohort of subjects among which 190 subjects were cognitively normal, 181 subjects had early MCI, 164 subjects had MCI, and 148 subjects had mild AD.

The ADNI-1 cohort was used in our initial analysis. To validate the findings of the initial analysis, a subset of the ADNI-2 cohort was used that included new subjects in the same disease categories (CN, MCI and AD) as the subjects in

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ADNI-1.

2.1.2 Imaging data

The neuroimaging phenotypes analyzed are predicted rates of change in cortical thickness and volumetric measurements in brain regions obtained from magnetic resonance imaging (MRI) scans. ADNI subjects had 1.5T MRI scans at either 6 or 12 month intervals during the two- to three-year follow-up period of the study and we chose to analyse the longitudinal information on cortical thickness and regional volumes. While other studies have compared the different study groups using imaging information from baseline (Shen et al., 2010; Meda et al., 2012), the longitudinal information provides insight into the different rates of brain deterioration experienced by people with negligible memory loss compared to those with more acute memory difficulties and Alzheimer's disease.

Segmentation Segmentation is the process of identifying the locations of anatomical structures within an image. In these data, the MRIs were segmented using Freesurfer (Fischl, 2012) software, identifying the locations of regions of interest such as the hippocampus, cerebellum and ventricles. For each hemisphere, the 28 volumetric and cortical thickness measurements used for analysis by Shen et al. (2010) were obtained via automated parcellation of the segmented images in Freesurfer. Cortical thickness (thickness of gray matter) and volumes of the regions of interest become increasingly atrophied as disease progresses, so we expect increased rates of atrophy in participants with more memory concerns (the MCI and AD subjects) compared to the cognitively normal subjects.

Predicting rates of change Linear mixed effect models, given in Equation 1, were used to predict the rates of change in each brain region of interest (ROI). Brief descriptions of the regions of interest are given in Appendix A. A separate mixed model was fit for each ROI, with random effects for subject-specific rates of change and fixed effects for average rates of change within diagnostic subgroups. The response variable Y_{ijt} is a continuous measurement of the cortical thickness or volume of a brain region of interest. In the specification of the model, fixed-effects terms are denoted by β, while random-effect terms are denoted by γ. The predictors are (i) t, the time of the follow-up visit at which the scan was conducted, with $t \in 0, 6, 12, 18, 24$ months; (ii) MCI, a dummy variable equal to 1 if subject i has mild cognitive impairment, and equal to 0 otherwise; and (iii) AD, a dummy

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

variable equal to 1 if subject i has Alzheimer's disease, and equal to 1 otherwise. The ROI's are indexed by j:

$$Y_{ijt} = \beta_{0j} + \beta_{1j}MCI + \beta_{2j}AD + \beta_{3j}t + \beta_{4j}MCI \times t + \beta_{5j}AD \times t + \gamma_{1ij} + \gamma_{2ij}t + \varepsilon_{ijt} \quad (1)$$

The predicted rate of change over the study period for subject i at ROI j is the sum of the disease-specific estimated rate of change and the subject-specific predicted rate of change $\hat{\beta}_{3j} + \hat{\beta}_{4j}MCI + \hat{\beta}_{5j}AD + \hat{\gamma}_{2ij}$. The lme4 package (Bates et al., 2015) was used to fit the mixed effects model in R, and the conditional mode was used to predict the random effect γ_{2ij}. The normality of the predicted, subject-specific, random effects was assessed by examining the Q-Q plots of the predictions at each ROI. The assumption of normality appeared to be reasonable for all ROIs.

Figure 2 is a heatmap of the predicted rates of change, adjusted for potential confounding variables in the sample, as discussed next. The heatmap illustrates how rates of change are more negative in subjects with more advanced disease, indicating that the thickness of gray matter and the volumes of brain regions of interest are shrinking more. Decreases in cortical thickness are more pronounced for subjects with AD (i.e. their rates of changes are more negative for various cortical thickness measures). Similarly, the ventricles, cavities in the brain filled with cerebrospinal fluid, have a more positive rate of change for subjects with more advanced disease. As brain atrophy progresses, these cavities expand.

2.1.3 Genomic data

The ADNI-1 subjects were genotyped with the Illumina Human610-Quad BeadChip and the ADNI-2 subjects were genotyped with the Illumina HumanOmniExpress BeadChip, both of which interrogate SNPs. All genotyping information was downloaded from the LONI Image Data Archive (Laboratory of Neuroimaging, 2015). Pre-packaged PLINK (Purcell et al., 2007) files included genotyping information for 757 of the 818 ADNI-1 subjects. Genotyping information for 793 of the ADNI-2 subjects were converted from CSV files to PLINK binary files using a publicly-available conversion script (Hibar, 2014). The Human610-Quad BeadChip and HumanOmniExpress BeadChip interrogated 620,901 and 730,525 SNPs respectively. APOE was genotyped separately at study screening from DNA extracted from a blood sample.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

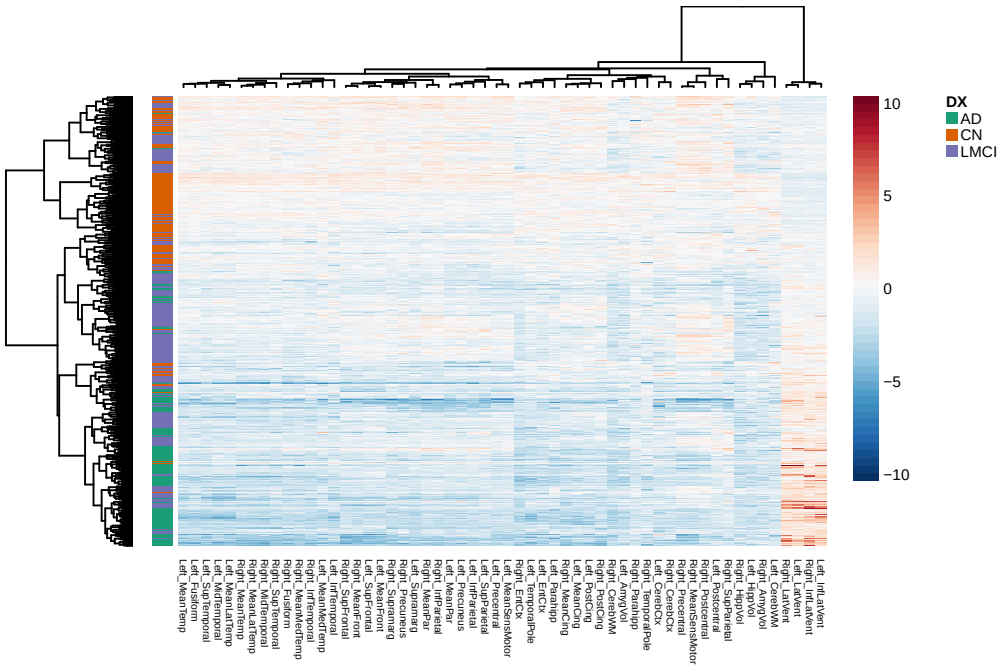

Figure 2: Heatmap of the neuroimaging phenotypes, clustered by similarity among ROIs and subjects. Each row corresponds to a subject in the sample and each column corresponds to one of the 56 ROIs. The rows are annotated by the disease group of the subject. The adjusted, predicted rates of change are shown for each region, where blue values indicate decreases in the volume of thickness in the brain region, and orange values indicate increases in the volume of the brain region. Values for the ventricles clustered on the far right, have an inverted relationships compared to the other ROIs since the ventricles are cavities in the brain which expand as brain atrophy progresses. The thickness of gray matter, by contrast, decreases with atrophy.

Inclusion criteria Subjects were included if their genotyping data were available, if they had a baseline MRI scan, and they had at least one additional follow-up baseline scan. Of the 757 ADNI-1 subjects and 793 ADNI-2 subjects with SNP data available, 696 and 583 had both a baseline scan and at least one additional follow-up scan, respectively. Genomic quality control procedures outlined below

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

in the Genomic quality control section were also applied, widening the exclusion criteria, to obtain a more homogeneous sample.

Genomic quality control Genomic quality control was performed with the freely-available software package PLINK. Thresholds for SNP and subject exclusion were as defined in Shen et al. (2010), except that we required a more conservative genotyping call rate for SNPs, as in Ge et al. (2012). Quality control was applied in three separate phases:

- Phase 1
    1. Exclude SNPs with genotyping call rate$<95\%$
    2. Restrict sample to subjects with self-reported non-Hispanic Caucasian ethnicity and white race.
- Phase 2
    1. Exclude SNPs with minor allele frequency (MAF) $<5\%$, Hardy-Weinberg equilibrium (HWE) $p<10^{-6}$.
    2. Exclude subjects based on tests for multivariate outliers and tests of relationship and gender using the genotyping data.
- Phase 3
    1. Exclude SNPs from sex chromosomes.

Genomic imputation Imputation serves two key roles in the analysis: to preserve the sample size for the multivariate analysis by replacing sporadically missing genotypes with imputed ones, and to impute SNPs not interrogated on the ADNI-2 chip that are interrogated on the ADNI-1 chip. Best-guess SNP genotypes were imputed in the ADNI-1 and ADNI-2 sample using the HapMap3 panel with NCBI build 36/hg18 using IMPUTE2 (Marchini and Howie, 2010), based on the imputation protocol in the IMPUTE2: 1000 Genomes Imputation Cookbook (Luan et al., 2014). Haplotypes were phased with SHAPEIT (Delaneau et al., 2013), and PLINK and SHAPEIT/IMPUTE2 file formats were converted with GTOOL (Freeman, 2007–2012). Out of the 503,450 SNPs that passed quality control in the ADNI-1 sample, sporadically missing genotypes were imputed at the 459,517 SNPs that were also in the reference panel. Out of the 574,730 SNPs that passed quality control in the ADNI-2 sample, sporadically missing genotypes

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were imputed at the 270,074 SNPs that were also on the ADNI-1 chip and in the reference panel. The remaining 189,443 SNPs that were not genotyped in the ADNI-2 sample, but were in both the ADNI-1 sample and the reference panel, were imputed into the sample. The genotyping rate in the imputed data for the ADNI-2 sample was 98.2%, prior to filtering out SNPs with an IMPUTE2 info metric < 0.5. The IMPUTE2 metric measures the reliability of imputed genotypes for a SNP, and takes a value of 1 when there is no genotype uncertainty in the sample. When the metric is less than 0.5, the sample mean of the posterior variance of imputed genotypes is at least half the variance that would be expected if alleles were sampled at random (Marchini and Howie, 2010).

The quality of the imputation for a SNP is reported based on the IMPUTE2 certainty measure, which is the average posterior probability of the best-guess genotypes in the sample. In the ADNI-1 data, the average certainty was 100% for all SNPs in the Alzgene linkage regions. In the ADNI-2 data, the certainties ranged from 69.1% to 100%, but 98.4% of the 31,301 imputed SNPs in the Alzgene linkage regions (defined below) had certainties of 90% or more. Due to the scale of the analysis, the imputed genotypes were treated as known. Ignoring genotype uncertainty is expected to lead to underestimates of the variance in the analysis of the ADNI-2 but not the ADNI-1 data.

Alzgene linkage regions To focus the analysis on regions that are likely to contain causal genetic variation, SNPs were included in the analysis if they fell in the linkage regions reported by approximate physical position on the Alzgene website (Bertram et al., 2007; Biomedical Research Forum, 2013). These linkage regions have been identified in meta-analyses of family-based studies of Alzheimer's disease (Hamshere et al., 2007; Butler et al., 2009). A total of 75,845 SNPs from nine chromosomes were included in the analysis from the ADNI-1 sample. Table 1 shows the number of SNPs in the ADNI-1 sample that fall in each linkage region. After filtering SNPs that had an IMPUTE2 info metric < 0.5, 75,818 SNPs remained in the ADNI-2 sample.

2.1.4 Adjustments for confounding

Covariate information cannot be explicitly included in SCCA, so both the imaging and genomic data are adjusted for confounding variables in advance. Potential confounders in the analysis are population stratification and APOE genotype. Population stratification is the phenomenon of systematic differences in allele frequencies in a subpopulation arising because of differences in ancestry, while the

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chromosome	Band	Mb	*N*
1	p31.1-q31.1	83-185	12005
3	q12.3-q25.31	103-173	10689
6	p21.1-q15	43-91	6785
7	pter-q21.11	0-78	13292
8	p22-p21.1	13-28	4149
9	p22.3-p13.3	20-35	2868
9	q21.31-q32	80-100	3483
10	p14-q24	10-100	15274
17	q24.3-qter	67-79	2319
19	p13.3-qter	8-54	4981

Table 1: The chromosome, band, and location on the Mb scale of the linkage regions of interest. *N* denotes the number of SNPs in the ADNI-1 data that fall in each linkage region.

$\varepsilon 4$ allele of APOE is the largest known genetic risk factor for Alzheimer's disease (Corder et al., 1993). Since true population structure is not observed, we adjust for it in the data using the top ten principal coordinates from multidimensional scaling. The top 10 principal coordinates account for 2.6% of the variability in the genotype data. We also restrict the analysis to the white non-Hispanic subjects and remove multivariate outliers identified in the top two principal coordinates as noted in the Genomic quality control section. We adjust for APOE genotype as a precautionary measure, since it can account for the population stratification in the data, over and above the principal components or principal coordinates (Lucotte et al., 1997).

Ten principal coordinates for each of the ADNI-1 and ADNI-2 datasets were obtained using ten-dimensional multi-dimensional scaling on the pairwise IBS distance matrix, computed with PLINK from 121,795 and 118,012 approximately uncorrelated SNPs from the SNPs that passed quality control filters. The SNP genotypes used to estimate the principal coordinates were from the complete imputed data. The number of principal coordinate dimensions was chosen to follow a similar protocol for adjustment for population stratification using principal components, in which ten axes of variation are suggested (Price et al., 2006).

The data for analysis were obtained by adjusting the minor allele counts and predicted rates of change of the brain ROIs for the ten principal coordinates, as well as for dummy variables for APOE genotype, using weighted ordinary least squares regression. The weights account for certain diagnostic subgroups being

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

over-represented in the sample relative to their population frequency. Details on the computation of the weights are presented in the next section. The residuals from each regression comprised the genomic (X) and neuroimaging (Y) features analyzed.

2.2 Methods

Figure 3 illustrates each step in the data analysis process, from deriving inverse probability weights to account for the biased sampling in ADNI-1 and ADNI-2, to discovering, refining and validating association between the adjusted minor allele counts at SNPs, which we call the genotypes, and the adjusted predicted rates of change at the brain regions of interest, which we call the neuroimaging phenotypes.

2.2.1 Inverse probability weights

To account for the biased sampling in the ADNI-1 and ADNI-2 case-control studies, we estimated inverse probability weights for each subject (Horvitz and Thompson, 1952). As subjects with early MCI were excluded from ADNI-1, we defined the target population to be non-Hispanic, white Americans and Canadians aged 55-90 years who are cognitively normal or have been diagnosed with late MCI or Alzheimer's disease.

The Alzheimer's Association reports that 5.2 million Americans had Alzheimer's disease in 2014 (Alzheimer's Association, 2014). Additionally, data from the US census in 2010 (U.S. Census Bureau, 2011) indicate that approximately 23% of the American population is over the age of 55 and that the total population is 308 million people. Based on this information, the approximate proportion of the American population aged 55-90 years with Alzheimer's disease is $p_{AD} = 7.5\%$, rounded to the nearest half percent. This calculation assumes that individuals aged 90 or more years and patients diagnosed with early MCI represent negligible proportions of the population. We used a late MCI prevalence estimate of $p_{MCI} = 5\%$ based on an urban study of people aged 65+ in New York (Manly et al., 2005), and assumed that the remaining $p_{CN} = 87.5\%$ of the population of interest is cognitively normal. A breakdown of the number of subjects used in the analysis by study is given in Table 2.

The inverse probability weights for each disease group, w_{DX} for ADNI-1 and w^*_{DX} for ADNI-2, are computed as the assumed prevalence of the disease in the target population divided by the number of subjects sampled from the disease

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

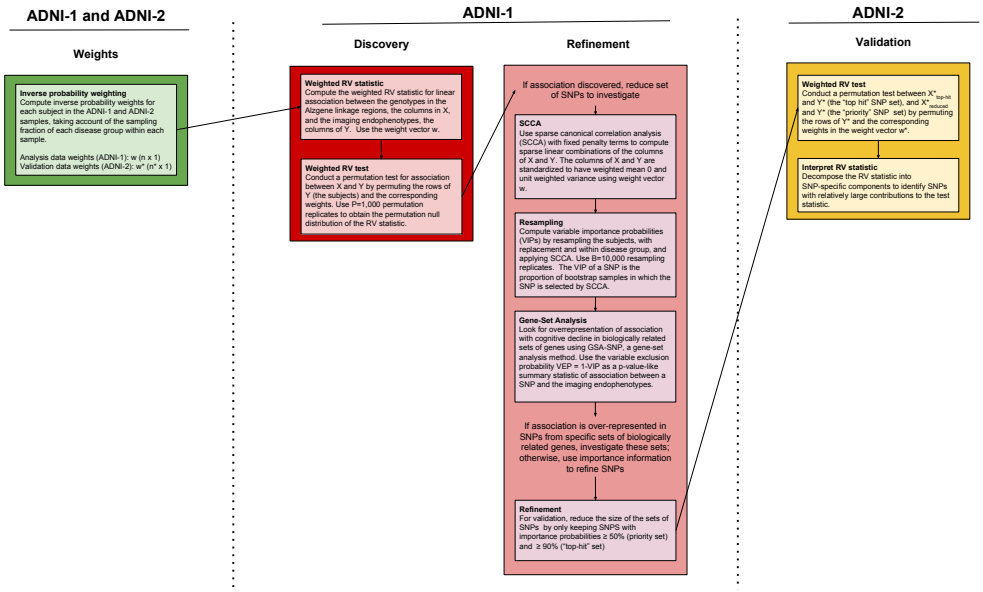

Figure 3: Flow chart of the analysis steps. The first row of column headings indicate the data sample used in the analysis step. The second row of column headings denote the step in the analysis beginning with computing weights, discovering association, refining the set of SNPs to investigate, and ending with validation of association with refined sets of SNPs and the neuroimaging phenotypes.

group. In the ADNI-1 sample, the MCI subjects have $w_{MCI} = 0.11$, AD subjects have $w_{AD} = 0.30$ and CN subjects have $w_{CN} = 3.09$, where the weights have been standardized to sum to the ADNI-1 sample size of $n = 632$. In the ADNI-2 sample, the MCI subjects have $w^*_{MCI} = 0.13$, the AD subjects have $w^*_{AD} = 0.44$ and the CN subjects have $w^*_{CN} = 2.00$, where the weights have been standardized to sum to the ADNI-2 sample size of $n = 265$.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Sample	n_{CN}	n_{MCI}	n_{AD}	n
ADNI-1	179	296	157	632
ADNI-2	116	104	45	265

Table 2: The number of subjects, n_D, from each disease group D that were analyzed in each study. The total number of subjects analyzed in each study is denoted by n.

2.2.2 Discovery

Weighted RV test We tested the analysis dataset, ADNI-1, for linear association between the genomic data and the neuroimaging data. The RV coefficient (Escoufier, 1973) is a multivariate generalization of Pearson's r^2 and quantifies the association between the columns of X, or the genotypes, and the columns of Y, the imaging endophenotypes. The coefficient can be defined in terms of the sample covariance matrices S_{XX} and S_{YY}, and the cross covariance matrix S_{XY} (Omelka and Hudecová, 2013), where the (k,l)th element of S_{XY} is defined as:

$$S_{X_k Y_l} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)(y_{il}-\bar{y}_l), \tag{2}$$

where $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$ and $\bar{y}_l = \frac{1}{n}\sum_{i=1}^{n} y_{il}$. The (k,l)th elements of S_{XX} and S_{YY} are defined similarly. The RV coefficient is then written as:

$$RV = \frac{\sum_k \sum_l S^2_{X_k Y_l}}{\sqrt{\sum_k \sum_l S^2_{X_k X_l} \sum_k \sum_l S^2_{Y_k Y_l}}}$$

and takes values between 0 and 1. When $RV = 0$, there is no linear association between the columns of X and Y, and higher values of RV are indicative of more association.

As the test statistic, we used a weighted version of the RV coefficient (Omelka and Hudecová, 2013) that calculates the sample variances and cross covariances accounting for the oversampling of AD and MCI patients in the study. A permutation test with P=10,000 permutations was used to assess the evidence for association between X and Y, where the rows of Y and their associated inverse probability weights were randomly permuted.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2.2.3 Refinement

SCCA and resampling Sparse canonical correlation analysis (SCCA; Parkhomenko et al., 2009) is a multivariate method for estimating maximally correlated sparse linear combinations of the columns of two multivariate datasets collected on the same n subjects, X and Y. To obtain a sparse linear combination of the SNP genotypes that is most associated with a non-sparse linear combination of the imaging phenotypes, we used SCCA, a penalized version of canonical correlation analysis. Sparse linear combinations contain some coefficients which are zero; e.g., in penalized regression analysis, the predicted value is potentially a sparse linear combination of the predictors. SCCA operates on the cross-correlation matrix, which is equivalent to the cross-covariance matrix S_{XY} as defined in Equation 2 when the columns of X and Y are standardized. SCCA estimates sparse linear combinations aX and bY that have maximal correlation, where a and b are column vectors of length p and q respectively. By operating on the cross-correlation matrix, the distance metric used is Euclidean, which is appropriate since both the columns of X and Y are treated as continuous in the analysis; other distance metrics for the genetic data are possible, however (Chang, 2017).

We initially applied SCCA to identify a sparse set of SNPs associated with the imaging endophenotypes. Ten-fold cross validation was used to select the penalty parameter for the SNPs, λ_u, for the SCCA. A search grid for λ_u was defined as $\{0, 10^{-4}, \ldots, 10^{-1}\}$ with the values in the search grid being incremented by 0.0005. At the i^{th} element in the search grid, $\lambda_{u,i}$, the sparse canonical-correlation coefficients, $a_{i,j}$ for the SNPs and $b_{i,j}$ for the endophenotypes, were computed in training set j. The fitted coefficients from the training sets were then used to compute the predicted sample-correlation coefficient in each test set: $\hat{r}_{i,j} = Corr(a_{i,j}X_{test_j}, b_{i,j}Y_{test_j})$. The SNP penalty parameter λ_u was chosen as the element in the search grid that maximized the sum of the predicted sample-correlation coefficients over the ten test sets. Under this cross-validation scheme, variable selection of the SNPs was minimal with more than 98% of the SNPs remaining in the active set. Ruling out fewer than 2% of the SNPs in the Alzgene linkage regions is insufficient refinement for our analysis.

Instead, we chose to incorporate bootstrap resampling to estimate the relative importance of each SNP in the multivariate association. This approach of "bootstrap enhancement" has been applied previously in neuroimaging studies (Bunea et al., 2011), to guide variable selection with the elastic-net and the lasso. We obtained B=100,000 bootstrap samples by sampling subjects with replacement

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

within each disease category. The weighted cross-correlation matrix $S_{XY}^{(W)}$ was computed for each bootstrap sample b, and a sparse linear combination of the genomic markers was estimated, using the SCCA penalty parameter $\lambda_u^* = 0.012$ for soft-thresholding the SNP coefficients. A value of $\lambda_u^* = 0.012$ was chosen so that approximately 10% of the SNPs had non-zero estimated coefficients. If $\beta_b = (\beta_{1b}, \beta_{2b}, \ldots, \beta_{pb})$ denotes the coefficient vector of the sparse linear combination of the p SNPs, from bootstrap sample b, then the importance probability for SNP k is defined as the proportion of bootstrap samples in which SNP k ($k = 1, \ldots, p$) has a nonzero coefficient, or is "selected":

$$VIP_k = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\hat{\beta}_{\mathrm{kb}} \neq 0), \text{ where } \mathbb{I}(\mathrm{A}) = 1 \text{ if condition } A \text{ holds and 0 otherwise}$$

Gene-Set Analysis To reduce the initial list of 75,845 SNPs to a shorter list for validation and to gain insight into the biologically related sets of genes associated with cognitive decline, we applied a gene-set analysis, as implemented in GSA-SNP (Nam et al., 2010). GSA-SNP combines the evidence for SNP-specific associations into gene-level summaries and assesses the pattern of association for genes in a given set, such as a functional pathway, relative to genes outside the set. We used variable exclusion probabilities, $VEP = 1 - VIP$, to quantify the SNP-specific evidence of association in lieu of p-values, and the second smallest VEP for SNPs in a gene as the gene-level summary statistic. Our choice of the second smallest was based on the recommendation of Nam et al. (2010) to use the second smallest rather than the smallest p-value as the gene-level summary statistic to protect against the spuriously high association signal that may be introduced by longer genes. The re-standardized version of GSA-SNP with the maxmean statistic (Efron and Tibshirani, 2007) was applied, with default gene padding of 20000 base pairs and Gene Ontology gene sets. We took $P = 100$ samples under the permutation null hypothesis of no association to serve as the empirical reference distribution for VEPs from ADNI-1. To ensure inclusive selection of SNPs, candidate gene sets were identified by Benjamini-Hochberg corrected p-values with a liberal threshold of 0.8 for the false discovery rate.

Refinement Two subsets of the SNPs in the Alzgene linkage regions, with estimated importance probabilities $\geq 50\%$ and 90%, were used for validation in the ADNI-2 sample. The cut-off values were chosen to reflect a relatively liberal and

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

stringent criterion, respectively. The set of SNPs with $VIP \geq 50\%$ is called the "priority set", while the SNPs with $VIP \geq 90\%$ is called the "top-hit" set.

2.2.4 Validation

Validation We assessed the evidence for linear association between the top-hit and priority SNPs and all the imaging phenotypes in the ADNI-2 validation sample using the RV-test with 1,000 permutation replicates.

Interpretation of RV statistic To further understand the observed association between SNPs in the priority set and endophenotypes in the ADNI-2 validation data, we decomposed the RV test statistic into its SNP-specific components. Details on how the SNP-specific contributions are calculated are described in the Results section.

3 Results

3.1 Discovery

The RV-test in the ADNI-1 data rejected the null hypothesis of no linear association between X and Y. The observed RV coefficient was $RV = 0.079$, and the permutation test p-value was $p < 0.001$.

3.2 Refinement

The resampling procedure coupled with SCCA in the ADNI-1 data produces variable importance probabilities (VIPs) for each SNP in the Alzgene linkage regions. Figure 4 is a Manhattan-like plot of the variable exclusion probabilities, $VEP = 1 - VIP$, plotted on the $-\log_{10}$ scale, such that SNPs with $VIP \geq 0.9$, have values of $-\log_{10}(VEP) \geq 1$. The dashed and dotted reference lines indicate the $VIP = 0.5$ and $VIP = 0.9$ cut-offs used to identify the priority and top-hit sets of SNPs, respectively.

1,694 SNPs had $VIP \geq 0.5$, a set of reduced SNPs we call the priority set. As expected, the priority SNPs, $X_{reduced}$, were associated with the endophenotypes in the ADNI-1 training data, based on a permutation RV test ($RV = 0.23$). Using the stringent cut-off of $VIP \geq 0.9$ for SNP selection, 22 SNPs were included in the

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

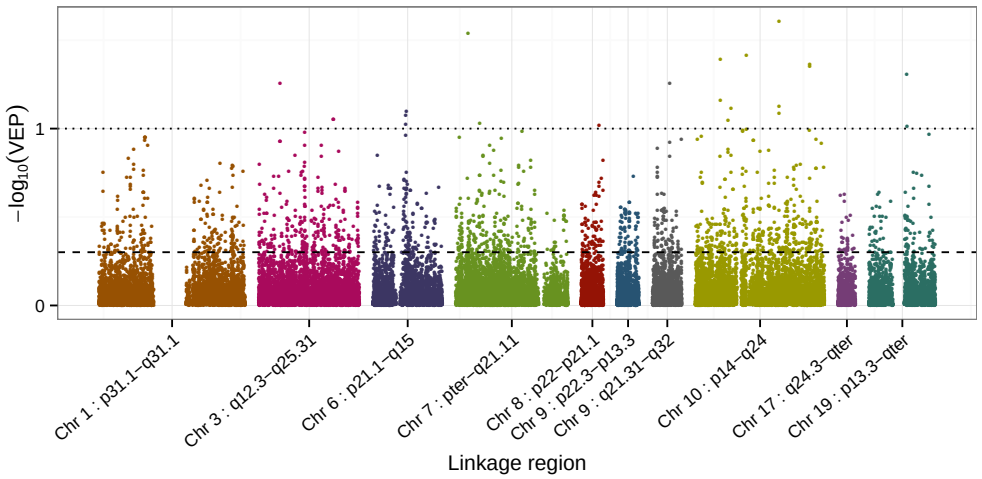

Figure 4: Plot of the $-\log_{10}(VEP)$ of the SNPs in each of the Alzgene linkage regions. The dashed and dotted reference lines indicate the $VIP = 0.5$ and $VIP = 0.9$ cut-offs used to define the priority and top-hit sets of SNPs, respectively.

"top-hit" set. There was no evidence of enrichment in biological pathways based on results from GSA-SNP.

Figure 4 shows that very few SNPs had $VIP \geq 0.9$, as evidenced by the sparse selection of SNPs with $-\log_{10}(VEP) \geq 1$ in the plot. While the linkage region on chromosome 10 is the largest, it also has the most SNPs with $VIP \geq 0.9$ and its SNPs have relatively high inclusion probabilities across the entire linkage region, in contrast to the linkage region from chromosome 6, for example. The smaller linkage regions p22.3-p-13.3 on chromosome 9 and q24.3-qter on chromosome 17 have relatively low inclusion probabilities, overall.

3.3 Validation

Let $X^*_{reduced}$ and X^*_{top} be the ADNI-2 validation data at the priority and top-hit sets of SNPs, respectively, and let Y^* be the validation endophenotype data. We were able to validate our finding of association between the priority set of SNPs and the endophenotypes in the ADNI-2 data. The RV test of association between $X^*_{reduced}$ and Y^* had an observed test statistic of $RV_{obs} = 0.073$, and a permutation

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

p-value of $p = 0.0021$. However, there was no evidence of association between the top-hit set of SNPs, X^*_{top}, and the endophenotypes ($p = 0.79$).

Figure 5 depicts the contribution of each SNP as a score, normalized to have mean 1 over all the SNPs in the priority set, to the RV statistic. Before normalization, the contribution for SNP i in the priority set is a sum, $\sum_{j=1}^{q} S^2_{X^*_{\text{reduced},i} Y^*_j}$, over the $q = 56$ endophenotypes in the cross-correlation matrix, $S^2_{X^*_{reduced} Y^*}$, from the ADNI-2 validation data. Each point in the plot therefore represents the relative contribution of a given SNP to the *RV* coefficient, summed over the 56 endophenotypes. SNPs with higher relative contributions can be viewed as the SNPs driving the association found in the *RV* test.

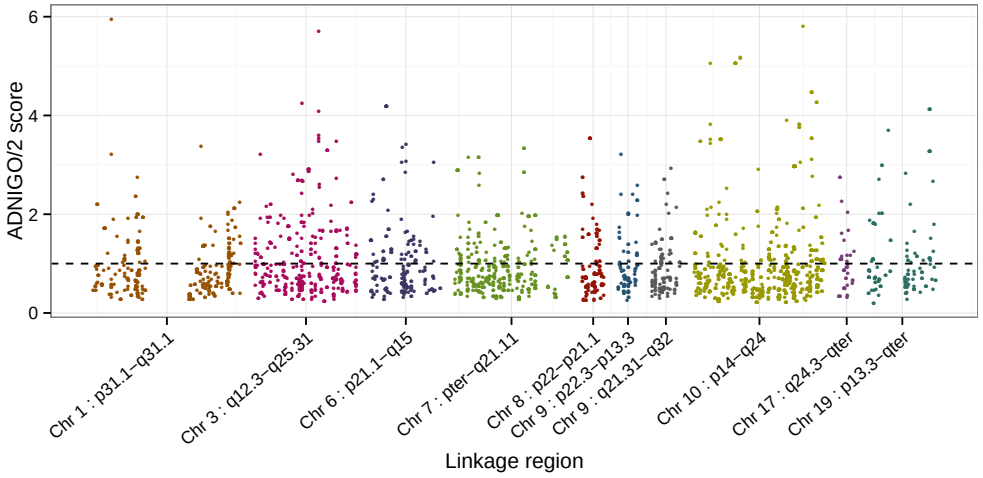

Figure 5: SNP-specific scores at the priority set SNPs in the ADNI-2 validation data, with scores defined as described in text. SNPs with higher score contribute relatively more to the *RV* coefficient between $X^*_{reduced}$ and Y^*. The dashed horizontal reference line corresponds to a score of 1, or the average score for a SNP in the priority set in the ADNI-2 validation data.

Table 3 summarizes information about the top 20 scoring SNPs in the priority set, with gene annotations obtained from SNPNexus (Ullah et al., 2012). SNPNexus was queried using assembly NCBI36/hg18, the UCSC genome browser

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(Speir et al., 2015) and AceView (Thierry-Mieg and Thierry-Mieg, 2006). The resulting gene symbols for annotated SNPs are reported in the Genes column of the table. We used the squared Pearson correlation coefficient, r^2, to measure the linkage disequilibrium (LD) between SNPs. Values of r^2 were computed in R with the snpMatrix package (Clayton and Leung, 2007) using the $N = 116$ cognitively normal subjects in the ADNI-2 data. LD blocks within the priority set are indicated by numbers in the first column of Table 3, and are defined such that all SNPs within a block have pairwise r^2 greater than 0.7.

LD block	SNP	score*	CHR	BP	Band	VIP	Genes
	rs17328231	5.96	1	95791119	1p31.1-q31.1	0.54	
	rs6439445	4.24	3	135119256	3q12.3-q25.31	0.72	
	rs16856619	4.08	3	146493435	3q12.3-q25.31	0.58	
	rs345015	5.70	3	146667792	3q12.3-q25.31	0.62	
1	rs634364	4.18	6	53575551	6p21.1-q15	0.57	AK126334, BC050580, AK125128, GCLC
1	rs525248	4.18	6	53576038	6p21.1-q15	0.57	AK126334, BC050580, AK125128
2	rs2148885	3.82	10	21413099	10p14-q24	0.57	NEBL
2	rs11012530	5.06	10	21444536	10p14-q24	0.58	NEBL
3	rs7897675	5.06	10	38448572	10p14-q24	0.55	ZNF37A
3	rs17588142	5.06	10	38467714	10p14-q24	0.55	
3	rs7080636	5.06	10	38659180	10p14-q24	0.50	
3	rs34350622	5.17	10	41848403	10p14-q24	0.51	
	rs12255371	5.15	10	41970728	10p14-q24	0.51	
	rs7088870	3.91	10	73723319	10p14-q24	0.59	
4	rs7094314	3.83	10	82321942	10p14-q24	0.55	SH2D4B
4	rs7904557	3.77	10	82326243	10p14-q24	0.56	SH2D4B
	rs12768174	5.82	10	84889167	10p14-q24	0.74	
	rs10887866	4.48	10	90661730	10p14-q24	0.56	STAMBPL1, KIAA1373, STAMBPL1andFAS
	rs4646957	4.26	10	94219892	10p14-q24	0.54	IDE
	rs1235382	4.13	19	49711347	19p13.3-qter	0.89	CEACAM20

* SNP-specific score indicating relative contribution to the RV statistic, as defined in text.

Table 3: The 20 SNPs with highest SNPs scores in the ADNI-2 dataset. Gene annotation obtained from SNPNexus queried with the UCSC genome browser and AceView. LD blocks comprise blocks of SNPs where all SNPs are in LD with $R^2 > 0.7$.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4 Discussion

In this report, we have taken a targeted approach to genetic association mapping of Alzheimer's disease by focusing on SNPs in Alzheimer's disease linkage regions and on imaging endophenotypes for brain regions affected by Alzheimer's disease. We discovered association between SNPs in the linkage regions and the imaging endophenotypes, refined the set of SNPs by selecting those with high variable inclusion probabilities, and validated the refined set in an independent dataset. Here, we discuss our observations about the benefits and pitfalls of applying data-integration methods such as sparse canonical correlation analysis and the RV test in a high-dimensional data setting with low signal. We also discuss potential links between Alzheimer's disease and genes in the priority set that were ranked highly in the validation data.

Initially, SCCA was used to find a subset of the SNPs in the linkage regions associated with the endophenotypes, but very little variable selection was achieved. SCCA uses a prediction criterion to identify the optimal soft-thresholding parameters for the sparse canonical variables, but using prediction error to select the penalty term includes irrelevant variables in the active set (Leng et al., 2006). In addition, the prediction-optimal value of the penalty term does not coincide with model selection consistency (Meinshausen and Bühlmann, 2006). Instead, we fixed the soft-thresholding parameter for the SNPs to achieve variable selection based on the rationale that no more than 7,500 SNPs (approximately 10%) are expected to be associated with the phenotypes. We applied bootstrapped-enhanced SCCA, a procedure analogous to the bootstrapped-enhanced elastic net proposed by Bunea et al. (2011) for imaging applications. To obtain a reduced set of SNPs to carry forward for validation, we thresholded the variable inclusion probabilities at 50%, as suggested by these authors, and at 90%. Bootstrapping to aid variable selection has been shown to be consistent in high-dimensional settings under some assumptions (Meinshausen and Bühlmann, 2010), and can improve recovery of the true model in regularized regression (Bach, 2008).

In a low-signal context, we do not necessarily expect to replicate association of the 22 "top-hit" SNPs in the validation data. For a fixed sample size, as the number of unassociated SNPs increases, the probability of a truly associated SNP being within the top-ranked SNPs decreases (Zaykin, 2005). By contrast, the more liberal threshold of $VIP \geq 50\%$ resulted in a larger, "priority" set of 1,694

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SNPs which could be validated and was substantially refined from the initial list of 75,845.

The permutation-based RV test of association proved to be a powerful tool in different phases of the analysis. This nonparametric test was computationally tractable and allowed us to uncover and validate linear association between the two multivariate datasets, one of them very high-dimensional, in an analysis setting with a low signal. Despite the evidence for association, the observed RV coefficient at each of the discovery, refinement and validation stages of the analysis was not large (< 0.1), consistent with SNPs having small association effects. The presence of SNPs with small effects is anticipated, as previous studies have found no large genetic effects apart from *APOE* (Ridge et al., 2013), for which we have already adjusted.

The ADNI studies use a case-control design, in which subjects are sampled conditional on meeting diagnostic criteria for either being cognitively normal, having late MCI, or having AD. Case-control designs do not result in a random sample from the population and they cannot be used to make inference about the population association between SNP genotypes and neuroimaging biomarkers without accounting for the biased sampling. To account for the biased sampling, we have applied inverse probability weighting in our analyses.

Investigation of the genes associated with the highest scoring SNPs in the validation data, reported in Table 3, identified genes previously implicated in AD. On chromosome 6, Glutamate-Cysteine Ligase Catalytic Subunit or *GCLC*, a gene annotation of the SNP rs634364, codes the first, rate-limiting enzyme of glutathione synthesis. Glutathione is an important antioxidant which plays an integrated role in the regulation of cell life, cell proliferation, and cell death (Pompella et al., 2003). The brain glutathione system is hypothesized to play a role in the breakdown of proteins in the brain, such as Aβ peptides (Lasierra-Cirujeda et al., 2013), and abundance of glutathione decreases with age and in some age-related disease (Liu et al., 2004). On chromosome 10, the complex locus *STAMPBL1andFAS* is an annotation of rs10887866 and codes a protein which plays a central role in programmed cell death (Choi and Benveniste, 2004). Through modulation of programmed cell death and neuronal atrophy, FAS may play a role in AD (Erten-Lyons et al., 2010). Also on chromosome 10, the gene insulin degrading enzyme (*IDE*) contains rs4646957 and codes the enzyme of the same name. *IDE* has pre-

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

viously been implicated in the progression Alzheimer's disease as it degrades the Aβ peptides which are the main components in the amyloid plaques on the brains of subjects with Alzheimer's disease (Edland et al., 2003). Edland et. al. found that three *IDE* variants were associated with risk of AD in subjects without copies of the ε4 *APOE* risk allele, the allele which constitutes the largest genetic risk of AD.

Gene expression from the UCSC RNA-Seq GTEx track was also explored to determine if any of the genes reported were highly expressed in the brain. On chromosome 10, Zinc Finger Protein 37A (*ZNF37A*), the gene containing rs7897675, is most highly expressed in the cerebellum and cerebellar hemisphere of the brain, regions related to motor function. Nebulette (*NEBL*), the gene annotation of rs2148885 and rs11012530, is most highly expressed in the heart, but has next highest gene expression in the brain. In addition, association fine-mapping under a linkage peak identified *NEBL* as a candidate gene for vitamin D levels in the blood (Aslibekyan et al., 2016). Low vitamin D blood levels are associated with accelerated decline in cognitive function in older adults (Miller et al., 2015).

Ten of the top 20 SNPs in Table 3 did not have associated gene annotations in the UCSC genome browser or AceView. For these SNPs, flanking genes were queried with ALFRED (Rajeevan et al., 2011) and the UCSC genome browser, since SNPs may "tag" causal variants in nearby genes. Genes were considered to be flanking if they were within 1 Mb of the SNPs in the priority set, though many of the flanking genes reported are much closer to the priority SNPs. On chromosome 3, rs643944 is approximately 22 kb proximal to the flanking gene *RAB6B*. *RAB6B* is the brain-specific isoform of *RAB6* (Wanschers et al., 2007), a family of proteins which impair the processing of the amyloid precursor protein involved in the development of AD (Thyrock et al., 2013). On chromosome 10, *DDIT4* is approximately 17.5 kb proximal to rs7088870. *DDIT4* produces the REDD1 protein, which enhances stress-dependent neuronal cell death and is involved in dysregulation of the mammalian target of rapamycin (mTOR) pathway (Maiese, 2014). Dysregulation of mTOR is a hallmark of a wide variety of brain disorders (Polman et al., 2012), and inhibition of mTOR is associated with Aβ-peptide-related synaptic dysfunction in AD (Ma et al., 2010). Another flanking gene to rs7088870 is *DNAJB12*, which is appoximately 39.2 kb proximal to rs7088870, and is involved in protein folding. The process of plaque build-up in AD involves the accumulation of misfolded Aβ proteins, and *DNAJB12* is

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

highly expressed throughout the brain (Tebbenkamp and Borchelt, 2010). Finally, in addition to being the gene annotation of rs7897674, *ZNF37A* is also 15.4 kb proximal to the SNP rs17588142 on chromosome 10.

In summary, this analysis illustrates the application of novel methods for integration of high-dimensional data with low signal. To focus on regions with increased prior probability of containing deleterious variants, the analysis was restricted to SNPs within linkage regions for AD. In practice, the same methodology could be applied to all available data if no linkage regions have previously been identified. However, we recommend restricting analyses to linkage regions whenever possible to leverage information from prior work. The objective was to obtain a refined list of SNPs to propose for further investigation. Naive application of SCCA did not lead to any refinement, potentially due to the data containing many small effects. Instead, we were able to obtain refinement through bootstrapped-enhanced SCCA. Throughout, the analysis benefited from the RV test to assess the evidence of linear association between two multivariate datasets: the high-dimensional genomic data, and the multidimensional neuroimaging data. RV tests of SNPs selected based on variable importance probabilities identified a priority set of 1,694 SNPs in the ADNI-1 data that was associated with the rates of changes in the brain regions of interest in the ADNI-2 validation set. Our final results are encouraging, in that genes corresponding to SNPs with the highest contributions to the RV coefficient in the validation data have previously been implicated or hypothesized to be implicated in AD, including *GCLC*, *IDE*, and *STAMBP1andFAS*. We hypothesize that the effect sizes of the 1,694 SNPs in the priority set are likely small, but further investigation within this set may advance understanding of the missing heritability in late-onset Alzheimer's disease.

5 Acknowledgements

Data collection and sharing was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun;

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

This work is based on Elena Szefer's MSc thesis supervised by J Graham and was supported in part by the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank Ellen Wisjman for helpful discussions about *APOE* and population stratification and Wayne Wang for assistance with the genomic quality control of the ADNI-2 validation data. The authors are also grateful to the anonymous reviewers whose feedback led to great improvements in the manuscript.

A

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ID	Measurement	Region of interest
AmygVol	Volume	Amygdala
CerebCtx	Volume	Cerebral cortex
CerebWM	Volume	Cerebral white matter
HippVol	Volume	Hippocampus
InfLatVent	Volume	Inferior lateral ventricle
LatVent	Volume	Lateral ventricle
EntCtx	Thickness	Entorhinal cortex
Fusiform	Thickness	Fusiform gyrus
InfParietal	Thickness	Inferior parietal gyrus
InfTemporal	Thickness	Inferior temporal gyrus
MidTemporal	Thickness	Middle temporal gyrus
Parahipp	Thickness	Parahippocampal gyrus
PostCing	Thickness	Posterior cingulate
Postcentral	Thickness	Postcentral gyrus
Precentral	Thickness	Precentral gyurs
Precuneus	Thickness	Precuneus
SupFrontal	Thickness	Superior frontal gyrus
SupParietal	Thickness	Superior parietal gyrus
SupTemporal	Thickness	Superior temporal gyrus
Supramarg	Thickness	Supramarginal gyrus
TemporalPole	Thickness	Temporal pole
MeanCing	Mean thickness	Caudal anterior cingulate, isthmus cingulate, posterior cingulate, and rostral anterior cingulate
MeanFront	Mean thickness	Caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole
MeanLatTemp	Mean thickness	Inferior temporal, middle temporal, and superior temporal gyri
MeanMedTemp	Mean thickness	Fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole
MeanPar	Mean thickness	Inferior and superior parietal gyri, supramarginal gyrus, and precuneus
MeanSensMotor	Mean thickness	Precentral and postcentral gyri
MeanTemp	Mean thickness	Inferior temporal, middle temporal, superior temporal, fusiform, parahippocampal, and lingual gyri, temporal pole and transverse temporal pole

Table 1A: Imaging phenotypes defined as volumetric or cortical thickness measures of $28 \times 2 = 56$ regions of interest (ROIs) from automated Freesurfer parcellations. Each of the phenotypes in the table corresponds to two phenotypes in the data: one for the left hemisphere and the other for the right hemisphere.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

Alzheimer's Association (2014): "2014 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, 10, 47–92, URL http://www.sciencedirect.com/science/article/pii/S1552526014000624.

Alzheimer's Disease Cooperative Study (2006): "ADNI procedures manual," URL http://adni-info.org/Scientists/doc/ADNI_GeneralProceduresManual.pdf.

Anna, S., L. Lena, F. Charlotte, K. Toru, W. Bengt, and G. Caroline (2011): "Linkage to the 8p21.1 Region Including the CLU Gene in Age at Onset Stratified Alzheimer's Disease Families," *Journal of Alzheimer's Disease*, 23, 1320, URL http://doi.org/10.3233/JAD-2010-101359.

Aslibekyan, S., L. K. Vaughan, H. W. Wiener, B. A. Hidalgo, D. J. Lemas, D. M. O'Brien, S. E. Hopkins, K. L. Stanhope, P. J. Havel, K. E. Thummel, B. B. Boyer, and H. K. Tiwari (2016): "Linkage and association analysis of circulating vitamin D and parathyroid hormone identifies novel loci in Alaska Native Yup'ik people," *Genes & Nutrition*, 11, –, URL http://dx.doi.org/10.1186/s12263-016-0538-y.

Bach, F. R. (2008): "Bolasso," in *Proceedings of the 25th international conference on Machine learning - ICML '08*, Association for Computing Machinery (ACM), URL http://dx.doi.org/10.1145/1390156.1390161.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015): "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67, URL http://doi.org/10.18637%2Fjss.v067.i01.

Beecham, G. W., E. R. Martin, Y.-J. Li, M. A. Slifer, J. R. Gilbert, J. L. Haines, and M. A. Pericak-Vance (2009): "Genome-wide Association Study Implicates a Chromosome 12 Risk Locus for Late-Onset Alzheimer Disease," *The American Journal of Human Genetics*, 84, 35–43, URL http://dx.doi.org/10.1016/j.ajhg.2008.12.008.

Bertram, L., C. Lange, K. Mullin, M. Parkinson, M. Hsiao, M. F. Hogan, B. M. Schjeide, B. Hooli, J. DiVito, I. Ionita, H. Jiang, N. Laird, T. Moscarillo, K. L. Ohlsen, K. Elliott, X. Wang, D. Hu-Lince, M. Ryder, A. Murphy, S. L. Wagner, D. Blacker, K. D. Becker, and R. E. Tanzi (2008): "Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Addition to APOE," *The American Journal of Human Genetics*, 83, 623–632, URL http://dx.doi.org/10.1016/j.ajhg.2008.10.008.

Bertram, L., M. B. McQueen, K. Mullin, D. Blacker, and R. E. Tanzi (2007): "Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database," *Nature Genetics*, 39, 17–23, URL http://dx.doi.org/10.1038/ng1934.

Biomedical Research Forum, L. (2013): "ALZGENE - PUTATIVE AD LINKAGE REGIONS BASED ON JOINT ANALYSES BY HAMSHERE ET AL. (2007), AND META-ANALYSES BY BUTLER ET AL. (2009)," URL http://www.alzgene.org/linkage.asp.

Bunea, F., Y. She, H. Ombao, A. Gongvatana, K. Devlin, and R. Cohen (2011): "Penalized least squares regression methods and applications to neuroimaging," *NeuroImage*, 55, 1519–1527, URL http://dx.doi.org/10.1016/j.neuroimage.2010.12.028.

Butler, A. W., M. Y. Ng, M. L. Hamshere, P. Forabosco, R. Wroe, A. Al-Chalabi, C. M. Lewis, and J. F. Powell (2009): "Meta-analysis of linkage studies for Alzheimer's disease—A web resource," *Neurobiology of Aging*, 30, 1037–1047, URL http://dx.doi.org/10.1016/j.neurobiolaging.2009.03.013.

Carbonetto, P. and M. Stephens (2012): "Scalable Variational Inference for Bayesian Variable Selection in Regression and Its Accuracy in Genetic Association Studies," *Bayesian Anal.*, 7, 73–108, URL http://dx.doi.org/10.1214/12-ba703.

Chalise, P. and B. L. Fridley (2012): "Comparison of penalty functions for sparse canonical correlation analysis," *Computational Statistics & Data Analysis*, 56, 245–254, URL http://dx.doi.org/10.1016/j.csda.2011.07.012.

Chang, C. (2017): "Distance matrices - PLINK 1.9," URL http://www.cog-genomics.org/plink/1.9/distance.

Choi, C. and E. N. Benveniste (2004): "Fas ligand/Fas system in the brain: regulator of immune and apoptotic responses," *Brain Research Reviews*, 44, 65–81, URL http://dx.doi.org/10.1016/j.brainresrev.2003.08.007.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Clayton, D. and H.-T. Leung (2007): "An R Package for Analysis of Whole-Genome Association Studies," *Human Heredity*, 64, 45–51, URL http://dx.doi.org/10.1159/000101422.

Corder, E., A. Saunders, W. Strittmatter, D. Schmechel, P. Gaskell, G. Small, A. Roses, J. Haines, and M. Pericak-Vance (1993): "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, 261, 921–923, URL http://dx.doi.org/10.1126/science.8346443.

Delaneau, O., J.-F. Zagury, and J. Marchini (2013): "Improved whole-chromosome phasing for disease and population genetic studies," *Nature Methods*, 10, 5–6.

Donoho, D. L. and J. M. Johnstone (1994): "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455, URL http://dx.doi.org/10.1093/biomet/81.3.425.

Edland, S., F. W.-D. Vriesé, D. Compton, G. Smith, R. Ivnik, B. Boeve, E. Tangalos, and R. Petersen (2003): "Insulin degrading enzyme (IDE) genetic variants and risk of Alzheimer's disease: evidence of effect modification by apolipoprotein E (APOE)," *Neuroscience Letters*, 345, 21–24, URL http://dx.doi.org/10.1016/s0304-3940(03)00488-9.

Efron, B. and R. Tibshirani (2007): "On testing the significance of sets of genes," *Ann. Appl. Stat.*, 1, 107–129, URL http://dx.doi.org/10.1214/07-aoas101.

Ertekin-Taner, N. (2003): "Fine mapping of the -T catenin gene to a quantitative trait locus on chromosome 10 in late-onset Alzheimer's disease pedigrees," *Human Molecular Genetics*, 12, 3133–3143, URL http://dx.doi.org/10.1093/hmg/ddg343.

Erten-Lyons, D., A. Jacobson, P. Kramer, A. Grupe, and J. Kaye (2010): "The FAS gene brain volume, and disease progression in Alzheimer's disease," *Alzheimer's & Dementia*, 6, 118–124, URL http://dx.doi.org/10.1016/j.jalz.2009.05.663.

Escoufier, Y. (1973): "Le Traitement des Variables Vectorielles," *Biometrics*, 29, 751, URL http://dx.doi.org/10.2307/2529140.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fallin, M. D., M. Szymanski, R. Wang, A. Gherman, S. S. Bassett, and D. Avramopoulos (2010): "Fine mapping of the chromosome 10q11-q21 linkage region in Alzheimer's disease cases and controls," *Neurogenetics*, 11, 335–348, URL http://dx.doi.org/10.1007/s10048-010-0234-9.

Fan, J., S. Guo, and N. Hao (2011): "Variance estimation using refitted cross-validation in ultrahigh dimensional regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 37–65, URL http://dx.doi.org/10.1111/j.1467-9868.2011.01005.x.

Fischl, B. (2012): "FreeSurfer," *NeuroImage*, 62, 774–781, URL http://dx.doi.org/10.1016/j.neuroimage.2012.01.021.

Freeman, C. (2007–2012): "GTOOL," http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html.

Gatz, M., C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, and N. L. Pedersen (2006): "Role of Genes and Environments for Explaining Alzheimer Disease," *Arch Gen Psychiatry*, 63, 168, URL http://dx.doi.org/10.1001/archpsyc.63.2.168.

Ge, T., J. Feng, D. P. Hibar, P. M. Thompson, and T. E. Nichols (2012): "Increasing power for voxel-wise genome-wide association studies: The random field theory least square kernel machines and fast permutation procedures," *NeuroImage*, 63, 858–873, URL http://dx.doi.org/10.1016/j.neuroimage.2012.07.012.

Greenlaw, K., E. Szefer, J. Graham, M. Lesperance, F. S. Nathoo, and A. D. N. Initiative (2017): "A bayesian group sparse multi-task regression model for imaging genetics," *Bioinformatics*, btx215, URL http://doi.org/10.1093/bioinformatics/btx215.

Hamshere, M. L., P. A. Holmans, D. Avramopoulos, S. S. Bassett, D. Blacker, L. Bertram, H. Wiener, N. Rochberg, R. E. Tanzi, A. Myers, F. W.-D. Vrieze, R. Go, D. Fallin, S. Lovestone, J. Hardy, A. Goate, M. O'Donovan, J. Williams, and M. J. Owen (2007): "Genome-wide linkage analysis of 723 affected relative pairs with late-onset Alzheimer's disease," *Human Molecular Genetics*, 16, 2703–2712, URL http://dx.doi.org/10.1093/hmg/ddm224.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Hibar, D. (2014): "ADNI_Genetics_Convert_to_PLINK," https://github.com/dhibar/ADNI_Genetics_Convert_to_PLINK, [commit: 10a49fe55b6bfd9746a1e027292660705e26a045].

Hoggart, C. J., J. C. Whittaker, M. D. Iorio, and D. J. Balding (2008): "Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies," *PLoS Genetics*, 4, e1000130, URL http://dx.doi.org/10.1371/journal.pgen.1000130.

Horvitz, D. G. and D. J. Thompson (1952): "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685, URL http://dx.doi.org/10.1080/01621459.1952.10483446.

Kamboh, M. I., F. Y. Demirci, X. Wang, R. L. Minster, M. M. Carrasquillo, V. S. Pankratz, S. G. Younkin, A. J. Saykin, G. Jun, C. Baldwin, M. W. Logue, J. Buros, L. Farrer, M. A. Pericak-Vance, J. L. Haines, R. A. Sweet, M. Ganguli, E. Feingold, S. T. DeKosky, O. L. Lopez, and M. M. Barmada (2012): "Genome-wide association study of Alzheimer's disease," *Translational Psychiatry*, 2, e117, URL http://dx.doi.org/10.1038/tp.2012.45.

Laboratory of Neuroimaging (2015): "LONI Image Data Archive," http://ida.loni.usc.edu, accessed: 2015-07-01.

Lasierra-Cirujeda, J., P. Coronel, M. Gimeno, and M. Aza (2013): "Beta-amyloidolysis and glutathione in Alzheimer's disease," *Journal of Blood Medicine*, 2013, 31, URL http://dx.doi.org/10.2147/jbm.s35496.

Leng, C., Y. Lin, and G. Wahba (2006): "A note on the lasso and related procedures in model selection," *Statistica Sinica*, 16, 1273–1284.

Liu, H., H. Wang, S. Shenvi, T. M. Hagen, and R.-M. Liu (2004): "Glutathione Metabolism during Aging and in Alzheimer Disease," *Annals of the New York Academy of Sciences*, 1019, 346–349, URL http://dx.doi.org/10.1196/annals.1297.059.

Luan, J., A. Teumer, J.-H. Zhao, C. Fuchsberger, and C. Willer (2014): "IMPUTE2: 1000 Genomes Imputation Cookbook," http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook, accessed: 2014-06-15.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Lucotte, G., F. Loirat, and S. Hazout (1997): "Pattern of gradient of apolipoprotein E allele *4 frequencies in western Europe." *Hum Biol*, 69, 253–62.

Lykou, A. and J. Whittaker (2010): "Sparse CCA using a Lasso with positivity constraints," *Computational Statistics & Data Analysis*, 54, 3144–3157, URL http://dx.doi.org/10.1016/j.csda.2009.08.002.

Ma, T., C. A. Hoeffer, E. Capetillo-Zarate, F. Yu, H. Wong, M. T. Lin, D. Tampellini, E. Klann, R. D. Blitzer, and G. K. Gouras (2010): "Dysregulation of the mTOR Pathway Mediates Impairment of Synaptic Plasticity in a Mouse Model of Alzheimer's Disease," *PLoS ONE*, 5, e12845, URL http://dx.doi.org/10.1371/journal.pone.0012845.

Maiese, K. (2014): "Taking aim at Alzheimer's disease through the mammalian target of rapamycin," *Annals of Medicine*, 46, 587–596, URL http://dx.doi.org/10.3109/07853890.2014.941921.

Manly, J. J., S. Bell-McGinty, M.-X. Tang, N. Schupf, Y. Stern, and R. Mayeux (2005): "Implementing Diagnostic Criteria and Estimating Frequency of Mild Cognitive Impairment in an Urban Community," *Arch Neurol*, 62, 1739, URL http://dx.doi.org/10.1001/archneur.62.11.1739.

Marchini, J. and B. Howie (2010): "Genotype imputation for genome-wide association studies," *Nat Rev Genet*, 11, 499–511, URL http://dx.doi.org/10.1038/nrg2796.

Meda, S. A., B. Narayanan, J. Liu, N. I. Perrone-Bizzozero, M. C. Stevens, V. D. Calhoun, D. C. Glahn, L. Shen, S. L. Risacher, A. J. Saykin, and G. D. Pearlson (2012): "A large scale multivariate parallel ICA method reveals novel imaging-genetic relationships for Alzheimer's disease in the ADNI cohort," *NeuroImage*, 60, 1608–1621, URL http://dx.doi.org/10.1016/j.neuroimage.2011.12.076.

Meinshausen, N. and P. Bühlmann (2006): "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, 34, 1436–1462, URL http://dx.doi.org/10.1214/009053606000000281.

Meinshausen, N. and P. Bühlmann (2010): "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72, 417–473, URL http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Miller, J. W., D. J. Harvey, L. A. Beckett, R. Green, S. T. Farias, B. R. Reed, J. M. Olichney, D. M. Mungas, and C. DeCarli (2015): "Vitamin D Status and Rates of Cognitive Decline in a Multiethnic Cohort of Older Adults," *JAMA Neurology*, 72, 1295, URL http://dx.doi.org/10.1001/jamaneurol.2015.2115.

Mueller, S. G., M. W. Weiner, L. J. Thal, R. C. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett (2005): "Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI)," *Alzheimer's & Dementia*, 1, 55–66, URL http://dx.doi.org/10.1016/j.jalz.2005.06.003.

Nam, D., J. Kim, S.-Y. Kim, and S. Kim (2010): "GSA-SNP: a general approach for gene set analysis of polymorphisms," *Nucleic Acids Research*, 38, W749–W754, URL http://dx.doi.org/10.1093/nar/gkq428.

Nathoo, F. S., K. Greenlaw, and M. Lesperance (2016): "Regularization parameter selection for a bayesian group sparse multi-task regression model with application to imaging genomics," in *Pattern Recognition in Neuroimaging (PRNI), 2016 International Workshop on*, IEEE, 1–4.

Nathoo, F. S., L. Kong, and H. Zhu (2017): "A Review of Statistical Methods in Imaging Genetics," *ArXiv e-prints*, URL http://arxiv.org/pdf/1707.07332.pdf.

Omelka, M. and Š. Hudecová (2013): "A comparison of the Mantel test with a generalised distance covariance test," *Environmetrics*, 24, 449–460, URL http://dx.doi.org/10.1002/env.2238.

Parkhomenko, E., D. Tritchler, and J. Beyene (2009): "Sparse Canonical Correlation Analysis with Application to Genomic Data Integration," *Statistical Applications in Genetics and Molecular Biology*, 8, 1–34, URL http://dx.doi.org/10.2202/1544-6115.1406.

Polman, J. A. E., R. G. Hunter, N. Speksnijder, J. M. van den Oever, O. B. Korobko, B. S. McEwen, E. R. de Kloet, and N. A. Datson (2012): "Glucocorticoids Modulate the mTOR Pathway in the Hippocampus: Differential Effects Depending on Stress History," *Endocrinology*, 153, 4317–4327, URL http://dx.doi.org/10.1210/en.2012-1255.

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Pompella, A., A. Visvikis, A. Paolicchi, V. D. Tata, and A. F. Casini (2003): "The changing faces of glutathione a cellular protagonist," *Biochemical Pharmacology*, 66, 1499–1503, URL http://dx.doi.org/10.1016/s0006-2952(03)00504-5.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006): "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, 38, 904–909, URL http://dx.doi.org/10.1038/ng1847.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham (2007): "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses," *The American Journal of Human Genetics*, 81, 559–575, URL http://dx.doi.org/10.1086/519795.

Radchenko, P. and G. M. James (2008): "Variable Inclusion and Shrinkage Algorithms," *Journal of the American Statistical Association*, 103, 1304–1315, URL http://dx.doi.org/10.1198/016214508000000481.

Rajeevan, H., U. Soundararajan, J. R. Kidd, A. J. Pakstis, and K. K. Kidd (2011): "ALFRED: an allele frequency resource for research and teaching," *Nucleic Acids Research*, 40, D1010–D1015, URL http://dx.doi.org/10.1093/nar/gkr924.

Ridge, P. G., S. Mukherjee, P. K. Crane, and J. S. K. Kauwe (2013): "Alzheimer's Disease: Analyzing the Missing Heritability," *PLoS ONE*, 8, e79771, URL http://dx.doi.org/10.1371/journal.pone.0079771.

Robert, P. and Y. Escoufier (1976): "A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient," *Applied Statistics*, 25, 257, URL http://dx.doi.org/10.2307/2347233.

Scott, W. K., J. M. Grubber, P. Conneally, G. W. Small, J. R. Gilbert, J. L. Haines, and M. A. Pericak-Vance (2000): "Fine-mapping of the chromosome 12 Alzheimer disease locus using family-based association tests of microsatellite markers," *Neurobiology of Aging*, 21, 129, URL http://dx.doi.org/10.1016/s0197-4580(00)82380-5.

Shen, L., S. Kim, S. L. Risacher, K. Nho, S. Swaminathan, J. D. West, T. Foroud, N. Pankratz, J. H. Moore, C. D. Sloan, M. J. Huentelman, D. W. Craig,

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

B. M. DeChairo, S. G. Potkin, C. R. Jack, M. W. Weiner, and A. J. Saykin (2010): "Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort," *NeuroImage*, 53, 1051–1063, URL http://dx.doi.org/10.1016/j.neuroimage.2010.01.042.

Speir, M. L., A. S. Zweig, K. R. Rosenbloom, B. J. Raney, B. Paten, P. Nejad, B. T. Lee, K. Learned, D. Karolchik, A. S. Hinrichs, S. Heitner, R. A. Harte, M. Haeussler, L. Guruvadoo, P. A. Fujita, C. Eisenhart, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn, and W. J. Kent (2015): "The UCSC Genome Browser database: 2016 update," URL http://dx.doi.org/10.1101/027037.

Tebbenkamp, A. T. N. and D. R. Borchelt (2010): "Analysis of Chaperone mRNA Expression in the Adult Mouse Brain by Meta Analysis of the Allen Brain Atlas," *PLoS ONE*, 5, e13675, URL http://dx.doi.org/10.1371/journal.pone.0013675.

Thierry-Mieg, D. and J. Thierry-Mieg (2006): "AceView: a comprehensive cDNA-supported gene and transcripts annotation," *Genome Biol*, 7, S12, URL http://dx.doi.org/10.1186/gb-2006-7-s1-s12.

Thyrock, A., E. Ossendorf, M. Stehling, M. Kail, T. Kurtz, G. Pohlentz, D. Waschbüsch, S. Eggert, E. Formstecher, J. Müthing, K. Dreisewerd, S. Kins, B. Goud, and A. Barnekow (2013): "A New Mint1 Isoform but Not the Conventional Mint1, Interacts with the Small GTPase Rab6," *PLoS ONE*, 8, e64149, URL http://dx.doi.org/10.1371/journal.pone.0064149.

Ullah, A. Z. D., N. R. Lemoine, and C. Chelala (2012): "SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update)," *Nucleic Acids Research*, 40, W65–W70, URL http://dx.doi.org/10.1093/nar/gks364.

U.S. Census Bureau (2011): "Table: Resident Population Data, 2010 Census," U.S. Department of Commerce, URL http://www.census.gov/2010census/data/apportionment-dens-text.php.

Wang, H., F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, and L. Shen (2011): "Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI

bioRxiv preprint doi: https://doi.org/10.1101/088310; this version posted July 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

cohort," *Bioinformatics*, 28, 229–237, URL http://dx.doi.org/10.1093/bioinformatics/btr649.

Wanschers, B. F., R. van de Vorstenbosch, M. A. Schlager, D. Splinter, A. Akhmanova, C. C. Hoogenraad, B. Wieringa, and J. A. Fransen (2007): "A role for the Rab6B Bicaudal–D1 interaction in retrograde transport in neuronal cells," *Experimental Cell Research*, 313, 3408–3420, URL http://dx.doi.org/10.1016/j.yexcr.2007.05.032.

Wechsler, D. (2009): *Wechsler Memory Scale: WMS-IV ; Technical and Interpretive Manual*, Pearson.

Witten, D. M., R. Tibshirani, and T. Hastie (2009): "A penalized matrix decomposition with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, 10, 515–534, URL http://dx.doi.org/10.1093/biostatistics/kxp008.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009): "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, 25, 714–721, URL http://dx.doi.org/10.1093/bioinformatics/btp041.

Zaykin, D. V. (2005): "Ranks of Genuine Associations in Whole-Genome Scans," *Genetics*, 171, 813–823, URL http://dx.doi.org/10.1534/genetics.105.044206.

Zhu, W., Y. Yuan, J. Zhang, F. Zhou, R. C. Knickmeyer, and H. Zhu (2016): "Genome-wide association analysis of secondary imaging phenotypes from the Alzheimer's disease neuroimaging initiative study," *NeuroImage*, In press, –, URL http://dx.doi.org/10.1016/j.neuroimage.2016.09.055.

Zou, H. and T. Hastie (2005): "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320, URL http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.

Züchner, S., J. R. Gilbert, E. R. Martin, C. R. Leon-Guerrero, P.-T. Xu, C. Browning, P. G. Bronson, P. Whitehead, D. E. Schmechel, J. L. Haines, and M. A. Pericak-Vance (2008): "Linkage and Association Study of Late-Onset Alzheimer Disease Families Linked to 9p21.3," *Annals of Human Genetics*, 72, 725–731, URL http://dx.doi.org/10.1111/j.1469-1809.2008.00474.x.