Multivariate genotypic analyses that identify specific genotypes to characterize disease and control groups in ADNI

Derek Beaton*

Rotman Research Institute

Baycrest Health Sciences

3560 Bathurst Street

Toronto, ON, Canada M6A 2E1

Jenny R. Rieck

Rotman Research Institute

Baycrest Health Sciences

3560 Bathurst Street

Toronto, ON, Canada M6A 2E1

Fahd Alhazmi

Department of Psychology

The Graduate Center,

City University of New York,

365 Fifth Ave, New York, NY 10016, USA

ADNI**

Alzheimer's Disease Neuroimaging Initiative

Hervé Abdi*

School of Behavioral and Brain Sciences

The University of Texas at Dallas MS: GR4.1,

800 West Campbell Rd.

Richardson, TX 75080 USA

Author Note

*Corresponding authors: dbeaton@research.baycrest.org, herve@utdallas.edu

**Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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 list of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

INTRODUCTION: Genetic contributions to Alzheimer's Disease (AD) are likely polygenic and not necessarily explained by uniformly applied linear and additive effects. In order to better understand the genetics of AD we require statistical techniques to address both polygenic and possible non-additive effects.

METHODS: We used partial least squares-correspondence analysis (PLS-CA)—a method designed to detect multivariate genotypic effects. We used ADNI-1 (N = 756) as a discovery sample with two forms of PLS-CA: diagnosis-based and ApoE-based. We used ADNI-2 (N = 791) as a validation sample with a diagnosis-based PLS-CA.

RESULTS: With PLS-CA we identified some expected genotypic effects (e.g., *APOE*/*TOMM40*, and *APP*) and a number of new effects that include, for examples, risk-associated genotypes in *RBFOX1* and *GPC6* and control-associated (possibly protective) genotypes in *PTPN14* and *CPNE5*.

DISCUSSION: To better understand the genetic contributions to AD we require techniques designed to detect complex (multivariate, genotypic) effects.

KEYWORDS*:* Genome-wide pattern analysis, partial least squares, correspondence analysis, genotypic model, discriminant analysis, statistical methods

1. Introduction

Many genes have been linked to Alzheimer's disease (AD) such as SORL1 [1], EPHA1 [2], CLU, and PICALM [3], all of which have shown effects in a variety of studies [4]. Though there are well-known genetic effects in AD [5–7]—there are still some controversial findings, for example, the work reported in [8] showed effects of rare variants in PLD3, where follow up studies showed both replication [9] and no replication [10] of PLD3 effects. Furthermore, recent works suggest that the genetic contributions to AD are: polygenic [11,12], epistatic [13], and non-linear or non-additive [14]. Therefore we must ask ourselves a methodological question: if the genetics of AD are so complex why are we still using methods whose assumptions are not suited to detect such complex effects?

With the advent of genome-wide studies, single nucleotide polymorphisms (SNPs) have almost exclusively been analyzed with the additive model. The additive model transforms a SNP from base pair letters into a count based on number of minor alleles where usually a major homozygote is "0," a heterozygote is "1," and a minor homozygote is "2." The additive model has become popular because it is viewed as both a practical approach [15] and a suitable model for complex traits [16]. However, the assumptions of additivity and linearity do not always hold (see [17,18]) and in some cases additive assumptions are detrimental because the *a priori* choice of a model that does not match the true inheritance pattern causes a loss of power [19]. Furthermore, the values of '0' and '2' are ambiguous across samples, and this ambiguity could lead to misinterpretation or dismissal of effects. For examples in AD: in [20] the authors reported "direction changes" in their replication analyses, whereas in [21] the authors report effects in the opposite direction in their attempt to replicate work in [8]. The additive model has been used in many AD studies even though the existence of non-additive effects are often acknowledged

[22,23]. With resources such as the Alzheimer's Disease Neuroimaging Initiative (ADNI), researchers have had the opportunity to analyze genome-wide data in a variety of ways, yet the ADNI genome-wide data have almost exclusively been analyzed with the additive model [24–29].

In this study we investigated the ADNI genome-wide data with a more general approach to avoid the potentially problematic assumptions of any particular model. To do so, we used partial least squares-correspondence analysis (PLS-CA, [30]). PLS-CA was designed to address the complexity of genetic contributions (i.e., polygenicity, non-additivity) and is a multivariate technique that detects rather than assumes genetic effects. Our study was designed to identify specific genotypic effects for AD (possible risk effects) and for controls (possible protective effects); this specificity helps disentangle some of the complexities of genetic contributions to AD.

Our paper is outlined as follows. In *Methods*, we describe the ADNI data used in this study, then describe the preprocessing and statistical techniques. We then detail the two phases of our study: (1) *Discovery* (with ADNI-1) which includes two genome-wide association analyses (one based on diagnosis, and another based on *APOE*), and (2) *Validation* (with ADNI-GO/2) that used the genotypes identified in *Discovery* to create a candidate panel and then test group association in a new sample. In *Results* we present findings from each of the phases. In *Discussion* we emphasize the *Validation* phase followed by a conclusions and limitations.

2. Methods

2.1 Data

Data used in the preparation of this article come from the ADNI database (http://adni.loni.usc.edu). The ADNI project was launched in 2003 as a public-private funding partnership and includes public funding by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and the Food and Drug Administration. The primary goal of ADNI has been to collect a wide variety of measures to assess the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations. Michael W. Weiner, MD (VA Medical Center and University of California at San Francisco) is the ADNI Principal Investigator. Subjects have been recruited from over 50 sites across the U.S. and Canada. For up-to-date information, see www.adni-info.org.

Our study included genomic, APOE, and diagnostic data from ADNI-1 and ADNI-GO/2. ADNI-1 used the Illumina Human610-Quad BeadChip while ADNI-GO/2 used the Illumina HumanOmniExpress BeadChip.

2.2 Participants

We obtained final totals of 756 participants from ADNI-1 (AD = 344, MCI = 204, CON = 208) and 791 participants from ADNI-GO/2 (AD = 203, MCI = 319, CON = 269). Table 1 includes demographics for ADNI-1 and ADNI-GO/2.

Table 1

Demographics of ADNI 1 and 2.

(a) Race (non-Hispanic/Latino)

	ADNI1	ADNI-GO/2
White	703 (685)	732 (700)
Asian	12 (12)	12 (12)
Black	37 (36)	30 (30)

American Indian/Alaskan	1 (1)	2 (2)
Hawaiian/Other Pacific Islander	0 (0)	2 (2)
Unknown/More than 1	3 (1)	13 (7)
Total	***756*** (***735***)	***791*** (***753***)

(b) Sex

	ADNI1	ADNI-GO/2
Female	308	367
Male	448	424
Total	***756***	***791***

(c) Diagnosis

	ADNI1	ADNI-GO/2
Control	208	269
Mild Cognitive Impairment	204	319
Alzheimer's Disease	344	203
Total	***756***	***791***

Note. Demographics information for ADNI 1 and ADNI-GO/2. All demographics information retrieved from the ADNIMERGE package. Diagnosis (Dx) was the last available diagnosis. Table (a) shows the distribution of race with the number of non-Hispanic/Latino in parentheses. "Unknown" and "More than 1" were combined into a single category. Table (b) shows the distribution of sex for each phase of ADNI. Table (c) shows the distribution of latest-available diagnosis (at the time of analyses) for each phase of ADNI.

2.3 Statistical techniques

Preprocessing, analyses, and graphics were performed primarily in `R` [31] and the `ExPosition`, `TExPosition`, and `TInPosition` packages [32,33]. Some in-house `MATLAB` (Mathworks Inc., Natick, MA) code was used for resampling.

Data were recoded into a disjunctive format (see Table 2). Because SNPs are categorical we required particular multivariate techniques designed specifically for categorical data. We used multiple correspondence analysis (MCA) instead of principal components analysis to correct for stratification effects.

Table 2

Nominal and disjunctive formats of data.

(a) Nominal

	SNP1 (with minor homozygote > 5%)	SNP2 (with minor homozygote < 5%)
Subj.1	AG	CA
Subj.2	AA	CA
…	…	…
Subj.i	AG	CC
...	…	…
Subj.I-1	<NA>	AA
Subj.I	GG	AA

(b) Disjunctive SNPs

	SNP1 (minor homozygote > 5%			SNP2 (minor homozygote < 5%)	
	AG	AA	GG	CA+CC	AA
Subj.1	1	0	0	1	0
Subj.2	0	1	0	1	0
…	…	…	…	…	…
Subj.i	1	0	0	0	0
…	…	…	…	…	…
Subj.I-1	.2	.7	.1	0	1
Subj.I	0	0	1	0	1

(c) Dx and ApoE

	Dx	#ApoE E4
Subj.1	AD	2
Subj.2	AD	1
…	…	…
Subj.i	MCI	1
...	…	…
Subj.I-1	MCI	2
Subj.I	CN	0

(d) Disjunctive Dx and ApoE

	Dx			# ApoE E4 Alleles		
	AD	MCI	CN	2	1	0
Subj.1	1	0	0	1	0	0
Subj.2	1	0	0	0	1	0
…	…	…	…	…	…	…
Subj.i	0	1	0	0	1	0
…	…	…	…	…	…	…
Subj.I-1	0	1	0	1	0	0
Subj.I	0	0	1	0	0	1

Note. Illustrative example of nominal (a and c) and disjunctive (b and d) coding of illustrative SNPs and diagnosis (Dx) plus ApoE. For SNP 1, all genotypes have a sufficient frequency and are coded (à la genotypic model), but for SNP 2, the minor homozygote (CC) does not occur frequently enough and is thus combined with the heterozygote (à la dominant model). In all tables, a 1 indicates the presence of a particular level of a categorical variable while a 0 indicates absence (e.g., Subj.1 is an Alzheimer's Disease patient, with 2 ApoE E4 alleles, the AG genotype for SNP1 and either a CC or CA [presence of minor allele] for SNP2). Note that one subject has missing data (i.e., "`<NA>`"). This subject's data for SNP 1 is imputed to the mean of the sample for that SNP (SNP 1) where AA occurs in 70% of the sample, AG in 20%, and GG in 10%, therefore the missing data are imputed to those values.

We used discriminant and seed PLS-CA in our discovery phase. Discriminant PLS-CA maximizes the separation between *a priori* groups of participants (see also [34]); we used Discriminant PLS-CA to identify genotypes most associated with each group. In Seed PLS-CA, a "seed" is a specific genetic marker, where the seed analysis looks for genes with distributions of genotypes similar to the seed (i.e., linkage disequilibrium); we used seed PLS-CA to identify genotypes with distributions similar to *APOE*.

2.4 SNP Preprocessing

For all analyses we excluded any SNPs in the X and Y chromosomes, the pseudoautosomal region (XY), and mitochondrial region (i.e., we analyzed SNPs in Chromosomes 1–22). All SNPs were preprocessed with `PLINK` (v1.07; [35]) and in-house `R` code. SNP annotation was performed with the `NCBI2R` package [36], and the `biomaRt` package [37,38]. SNPs were pre-processed with pipeline described below.

Participant and SNP call rates (i.e., completeness of data) ≥ 90%, minor allele frequency ≥ 5%, Hardy-Weinberg equilibrium $p \leq 10^{-6}$. SNPs were then recoded into a disjunctive format. Additionally, any genotype that fell below a 5% threshold was combined with another genotype. Only minor homozygotes were below the 5% threshold, and thus combined with the

heterozygotes—a procedure which is analogous to a dominant model. Missing genotypes were imputed to the mean of the sample (see Table 2).

2.5 Study design and overview

We conducted a two-part study: *Discovery* and *Validation*. In the *Discovery* phase there were two analyses with ADNI-1 genome-wide SNPs. The results from the *Discovery* analyses in ADNI-1 were used to create candidate SNP panels for validation in ADNI-GO/2. In the validation phase there was one analysis with a specific subset of ADNI-GO/2 SNPs.

Data from ADNI-1 and ADNI-GO/2 were not combined or preprocessed together at any stage in this study, so that no contamination or influence occurred from one set on the other in order to guarantee statistical independence for the discovery-validation pipeline. Because ADNI-1 and ADNI-GO/2 have two different chip sets, we generated a candidate panel of SNPs for ADNI-GO/2 based on the SNPs and their associated genes identified in ADNI-1 (*Discovery*).

2.6 Discovery study preprocessing and analyses

The goal of the Dx × genotype analyses was to detect genotypes most associated with each diagnostic category. The seed PLS-CA used *APOE* E4 as the seed, and thus was performed on *APOE* (0, 1, or 2 E4 alleles) × genotypes. The goal of the *APOE* × genotype analyses was to identify new candidate genotypes with a distribution similar to *APOE*. Both analyses should identify candidate markers of AD: the discriminant analysis (henceforth referred to as "Dx-GWAS") identifies genotypes most associated with each group, whereas the seed analysis (henceforth referred to as "ApoE-GWAS") identifies genotypes similar to *APOE*. All analyses used bootstrap resampling [39] to identify stable genotypes. The distributions around the genotype were tested with "bootstrap ratios" (BSR; [30]). Significant genotypes in our two

GWAS (in ADNI-1) were then used to create a new candidate panel of SNPs for validation (in ADNI-GO/2).

2.7 Creation of SNP panels from Discovery for Validation

Because ADNI-1 and ADNI-GO/2 were used as independent data sets in our study, and because the data come from two different genome-wide chips, we used the significant genotypes from the discovery analyses (i.e., Dx-GWAS and the ApoE-GWAS) to generate candidate SNPs for validation. We did so with the following procedure. For all significant genotypes in the discovery analyses we used their respective SNPs to: (1) compile a list of all SNPs within a 50kbase (25+/-) window of those SNPs, and (2) retrieve all stable `ensembl` gene (ENSG) IDs associated with those SNPs from the discovery analyses and in turn retrieved *all possible SNPs* associated those ENSG IDs. All SNPs from the steps (1) and (2) were aggregated. We extracted all SNPs from the ADNI-GO/2 data that were from the discovery-derived aggregate set for use in the validation phase.

2.8 Validation study preprocessing and analyses

Validation analyses were conducted on the ADNI-GO/2 data set with a discriminant PLS-CA on the validation SNPs based on diagnosis. The goal of the validation Dx × genotype analysis was to detect genotypes most associated with each diagnostic category. As in the discovery phase, we used bootstrap resampling and the BSR test to identify stable genotypes.

3. Results

3.1 Discovery (ADNI-1)

ADNI-1 genome-wide data contains 620,901 SNPs and 757 participants. After QC and preprocessing 756 participants (AD = 344, MCI = 204, CON = 208) and 517,441 SNPs (in chromosomes 1–22) remained, which produced a 756 participants × 1,332,455

disjunctive genotypes matrix (see Table 2). MCA was applied to the disjunctive genotypes matrix to identify confounding effects. The first two components showed effects of race and ethnicity but not sex and were thus removed from (i.e., regressed out of) the data. Subsequent components showed no apparent effects of stratification. For the discovery GWAS, we used a cutoff of ±5 for the BSR tests which is slightly below the traditional GWAS parametric threshold ($p < .05 \times 10^{-8}$ would correspond to a BSR ≈ 5.33).

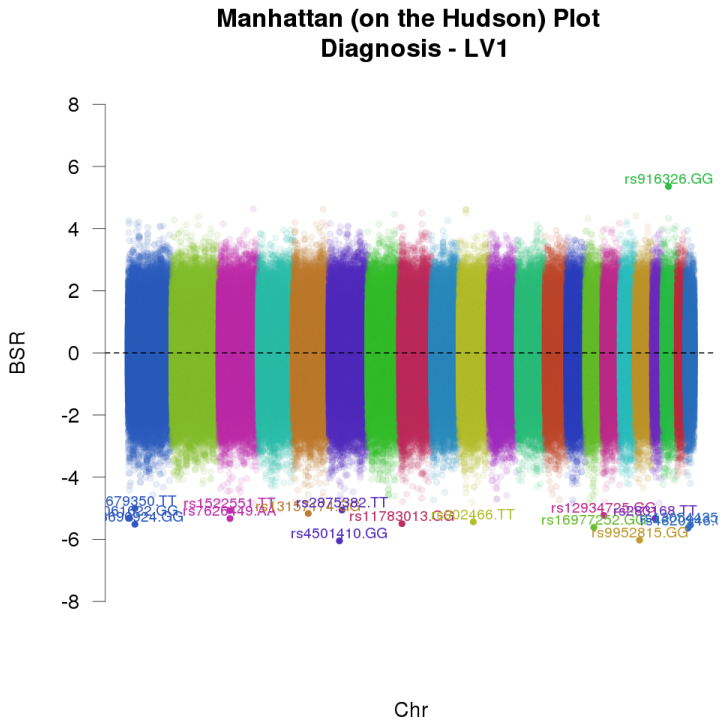
3.1.1 Dx-GWAS

The discriminant PLS-CA produced two components. Component 1 explained 50.25% of the variance and was driven by the separation of the AD group from the MCI group (see Supplemental Figure 1). Component 2 explained 49.76% of the variance and separated the CON group from the other two groups (see Supplemental Figure 1). Figure 1 shows all genotypes plotted with their BSR values in a Manhattan-like plot. BSR values can be positive or negative—because the sign matches their component score—we call this plot "Manhattan on the Hudson" (MotH; like a city skyline and reflection on a river). Figure 1a shows the BSRs for all genotypes for Dx-GWAS Component 1 and Figure 1b shows the BSRs for all genotypes for Dx-GWAS Component 2. The majority of significant genotypes were more related to the AD group than the other groups (see Figure S1 and Table S1).

3.1.2 ApoE-GWAS

Because there were only 3 levels to the seed ("0 E4," "1 E4," and "2 E4" alleles), seed PLS-CA produced only two components. Component 1 explained 51.24% of the total variance and was driven by the presence (left) vs. absence (right) of E4 alleles (see Supplemental Figure 2). Component 2 explained 48.76% of the variance and separates the two E4 alleles group from the other two groups (see Supplemental Figure 2). Figure 2 shows the BSRs for all genotypes

from this analysis in a MotH plot. Figure 2a shows the BSRs for all genotypes for ApoE-GWAS Component 1 and Figure 2b shows the BSRs for all genotypes for ApoE-GWAS Component 2 (see also Figure S2 and Table S2).



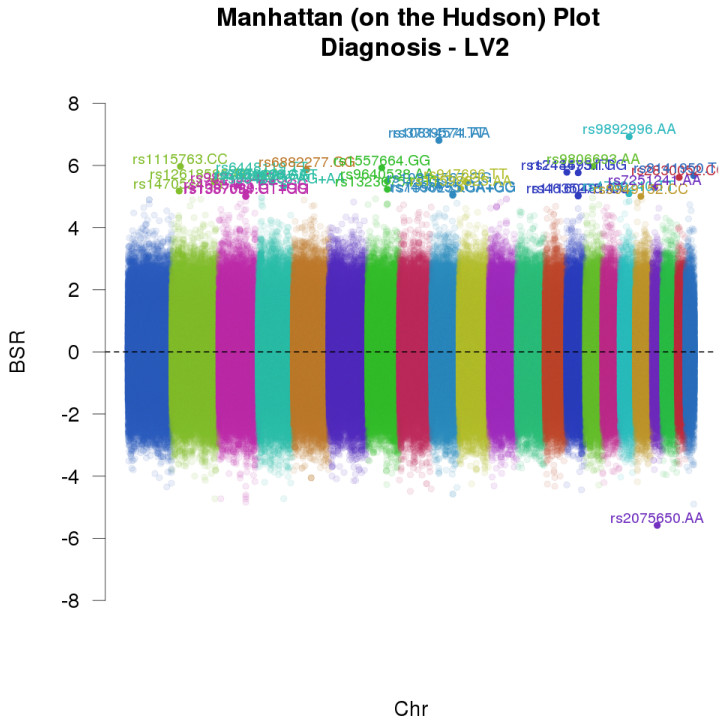
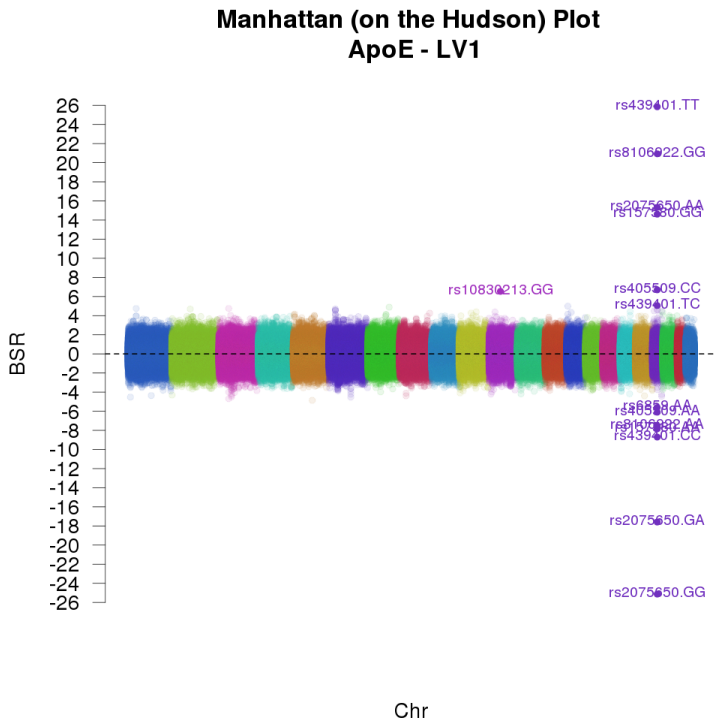


Figure 1. Manhattan (on the Hudson) plots for the multivariate Diagnosis (Dx)-GWAS. Horizontal axes are each genotype ordered by Chromosome (Chr), where each Chr is color-coded (1–22). The vertical axes are bootstrap ratio values (BSRs). Both panels show BSRs (analogous to *t*- or *Z*-scores; which can be positive or negative) for each genotype along Component 1 (a.k.a. Latent Variable (LV) 1; panel a, top) and Component 2 (a.k.a. LV 2; panel b, bottom) – the same components as in Figure 2. With respect to the multivariate Dx-GWAS a wide variety of genotypes show significant, and similar, effects and are not concentrated in any particular region (see also Table S1 and Figure S1).



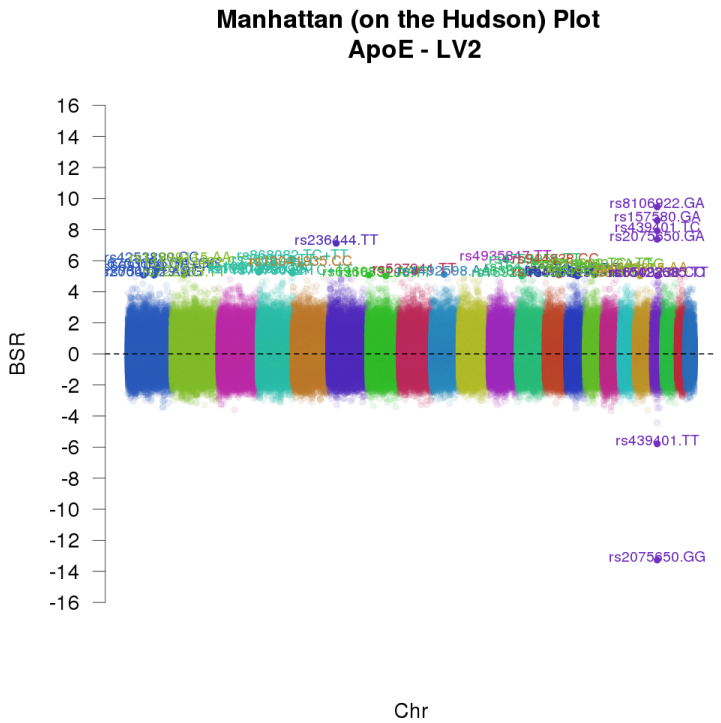


Figure 2. Manhattan (on the Hudson) plots for the multivariate ApoE-GWAS. Horizontal axes are each genotype ordered by Chromosome (Chr), where each Chr is color-coded (1–22). The vertical axis is bootstrap ratio values (BSRs). Both panels show BSRs (analogous to *t*- or *Z*-scores; which can be positive or negative) for each genotype along Component 1 (a.k.a. Latent Variable (LV) 1; panel a, left) and Component 2 (a.k.a. LV 2; panel b, right) – the same components as in Figure 4. With respect to the multivariate ApoE-GWAS, many of the effects are concentrated, generally, in Chr19 (see also Table S2) across both components, but much more so for Component 1 (a; left). While Component 2 (b; right) shows a variety of effects, some of the strongest are still in Chr19.

3.2 Candidate panel creation

In order to assess and understand the impact of the genotypes and their respective genes identified in *Discovery*, we require a *Validation* phase of independent data. All SNPs associated with significant genotypes in *Discovery* were used to create the candidate panel. From 96 SNPs (see Tables S1 and S2), our candidate process identified 1,045,360 possible SNPs; we extracted all available SNPs in ADNI-GO/2 for use in *Validation*.

3.3 Validation (ADNI-GO/2)

ADNI-GO/2 genome-wide data contains 730,525 SNPs and 791 participants. From the candidate panel, we extracted 5,508 SNPs. After QC and preprocessing, 791 participants (AD =

203, MCI = 319, CON = 269) and 5,508 SNPs remained, which produced a 791 participants × 14,200 disjunctive genotypes matrix (see Table 2). MCA was applied to the disjunctive genotypes matrix to identify confounding effects. The first two components showed effects of race and ethnicity but not sex and were thus removed from the data. Subsequent components showed no apparent effects of stratification.

A discriminant PLS-CA (Dx × genotypes) was performed on the 791 × 14,200 matrix. For the validation analysis, we used a cutoff of ±3.25 for the BSR tests (roughly equivalent to a Bonferroni cutoff for the number of unique ENSG IDs). Discriminant PLS-CA produced two components: Component 1 explained 50.41% and was driven by the separation of the MCI group from the AD and CON groups (see Figure 3); Component 2 explained 49.59% of the variance and separated the CON group from the AD group (see Figure 3).

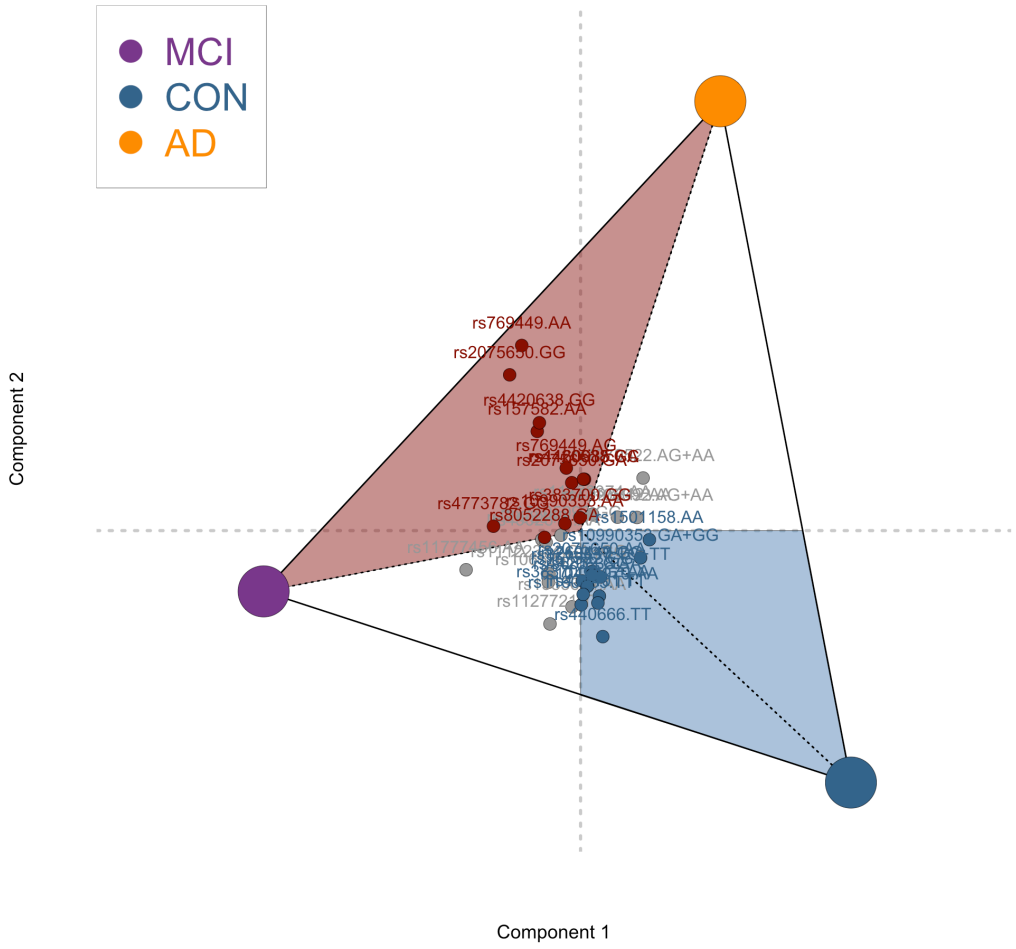


Figure 3. Diagnosis-based analysis in the *Validation* phase. The groups form a boundary region on the components. We denoted portions of this subspace as "control" associated (blue) or "risk" associated (red); anything outside the control or risk regions is ambiguous. Genotypes that fall in the "control" region are more associated with the control group than the other groups and genotypes that fall in the "risk" region are more associated with {MCI or AD} than the CON group.

Table 3

Validation set: significant genotypes on either component

rsid	genotype	BSR - 1	BSR - 2	chr	Gene symbol	Risk / Control Region
rs10494979	AA	0.95	*-3.422*	1	PTPN14	CON
rs11122374	AA	1.985	*3.255*	1	TSNAX-DISC1,DISC1	
rs11122374	GA+GG	-1.995	*-3.346*	1	TSNAX-DISC1,DISC1	
rs1501158	AA	*3.319*	-0.409	4	LINC00504	CON
rs1065261	TT	-1.971	*-3.451*	6	CPNE5	

rs3213537	TC+TT	1.345	*-3.31*	6	CPNE5	CON
rs11777456	AA	*-3.366*	-1.615	8	NCALD	
rs10990353	AA	*-3.374*	1.483	9	ZYG11AP1 / LOC100421294	RISK
rs10990353	GA+GG	*3.496*	-1.482	9	ZYG11AP1 / LOC100421294	CON
rs7093342	GA	*-3.395*	-0.848	10	ITIH5	
rs7093342	AA	*3.904*	1.5	10	ITIH5	
rs1979522	AG+AA	*3.275*	2.298	12	LRMP	
rs4773782	GG	*-3.608*	0.185	13	GPC6	RISK
rs17105992	GG	*-3.513*	-0.847	14	LOC107984016,RAD51B	
rs17105992	AG+AA	*3.55*	0.844	14	LOC107984016,RAD51B	
rs4902611	AA	*-3.93*	-1.293	14	LOC107984016,RAD51B	
rs8052288	GA	*-3.476*	-0.692	16	RBFOX1	RISK
rs1553614	AA	-0.278	*-3.272*	16	RBFOX1	
rs6859	GG	0.415	*-3.895*	19	NECTIN2‡	CON
rs2075650	GA	-0.656	*3.5*	19	TOMM40	RISK
rs2075650	AA	1.555	*-5.558*	19	TOMM40	CON
rs2075650	GG	-2.12	*4.338*	19	TOMM40	RISK
rs157582	GG	0.162	*-5.297*	19	TOMM40	CON
rs157582	AA	-1.773	*3.72*	19	TOMM40	RISK
rs1160985	CC	0.344	*4.349*	19	TOMM40	RISK
rs1160985	TT	0.048	*-4.101*	19	TOMM40	CON
rs769449	GG	1.749	*-6.527*	19	APOE	CON
rs769449	AA	-1.792	*5.195*	19	APOE	RISK
rs769449	AG	-0.995	*4.194*	19	APOE	RISK
rs439401	TT	0.678	*-3.277*	19	APOE	CON
rs4420638	GA	0.207	*3.95*	19	APOC1	RISK
rs4420638	AA	0.803	*-6.458*	19	APOC1	CON
rs4420638	GG	-1.637	*3.913*	19	APOC1	RISK
rs383700	GG	-0.179	*3.314*	21	APP	RISK
rs383700	AG+AA	0.179	*-3.436*	21	APP	CON
rs440666	TT	0.631	*-3.88*	21	APP	CON
rs1127721	TT	-0.789	*-3.543*	22	PARVG	

Note. ‡ rs6859 previously associated with PVRL2. Significant genotypes for both components from the validation set. Component 1 primarily separated CON from AD, whereas Component 2 primarily separated AD from the other two groups. In the validation analysis, we used the asymmetric version of correspondence analysis (CA) to help visualize which genotypes are most associated with each group, and to create identify “control” (CON) and “risk” (RISK) regions within the CA components (see Figure 3). Genotypes labeled as CON or RISK indicate which region they are in, otherwise their association is more ambiguous in terms of association with risk or control (“protective”) factors.

Table 3 lists all significant genotypes from *Validation*. In Figure 3, we highlight boundaries to interpret genotypic effects. We focused specifically on two boundaries: (1) the boundary associated with the CON group (lower right, blue) and (2) the boundary associated with "risk status" (i.e., AD or MCI; upper middle to middle left, red). The relevant SNPs are highlighted in their boundary colors in Figure 3, and identified as part of the "RISK" or "CON" region in Table 3.

4. Discussion

There has been an increased interest in multivariate approaches for the genetics of AD [40]. However studies in AD still only use the additive model despite the fact that more general models are better suited for complex effects. Toward this end, our study provided a unique approach that identifies more specific genetic contributions to AD than other studies.

4.1 Chromosome 19

Some of the strongest effects in all of our analyses were associated with genotypes in Chromosome 19 (Chr19); these effects include *TOMM40*, *APOE, APOC*, and *NECTIN2*, which are typically well known. However, PLS-CA identified exactly which genotypes contributed to which effects (see Fig. 3 and Table 3). For examples, the 'AA' genotype of rs769449 (*APOE*) and the 'GG' genotype of rs2075650 (*TOMM40*) are at the extreme of our plot almost exactly half way between AD and MCI and directly opposite of the CON group (Fig. 3). This means that 'AA' in rs769449 and GG in rs2075650 rarely occur in CON, but tend to occur roughly equally in both AD and MCI. Most importantly for Chr19 effects: our *Discovery* phase identified 'AA' of rs6859 associated with the *presence* of *APOE* E4, whereas our *Validation* phase identified 'GG' of rs6859 associated with the CON group. While the "direction" of the effect is the same

across both studies, the source of the effect was not (i.e., minor allele vs. major allele).

4.2 APP

APP is well-known in AD [41]. In our study, *APP* was identified in the Dx-GWAS in *Discovery* and also showed effects in the *Validation* analyses (see Table 3). We saw *APP* markers associated both with "risk" and "control" (Fig. 3 and Table 3). Given the findings across both our *Discovery* and *Validation* phases, our findings here suggest that *APP* is more related to diagnostic criteria than to *APOE* distribution, and that specific *APP* genotypes are possibly protective (e.g., 'TT' in rs440666).

4.3 RBFOX1 and GPC6.

We found risk specific genotypes in the *GPC6* and *RBFOX1* genes (Table 3; Fig. 3). *RBFOX1* has been associated with neurofibrillary tangles in AD [42] as well as imaging phenotypes [43] and hippocampal volumes [44] in AD. *GPC6* has been associated with cognitive decline in AD [45].

4.4 Novel genetic risk and "protective" factors.

Our analyses revealed several novel genetic "control" associated contributions: *PTPN14*, *CPNE5*, *LINC00504*. Furthermore, *ZYG11AP1* and *LOC100421294* each had various genotypes associated with "control" and "risk" effects. There are no reported effects of rs1501158 (*LINC00504*) and rs10990353 (*ZYG11AP1* / *LOC100421294*) in AD, dementia, or the cognitive aging literature. *CPNE5* has not been reported in AD but has been associated with other neurodegenerative diseases [46].

4.5 Ambiguous but stable effects.

In the validation analysis, there were "ambiguous" effects that we cannot classify as "risk" or "control" because they fall outside of our designated region. However these effects are of interest because (1) these gene effects appear in both analyses and (2) of their associations in various disorders or their interactions with other genes. *PARVG* has been associated with Parkinson's Disease [47] and neurodegeneration [48]. Furthermore, *RAD51B*, *NCALD*, and *DISC1* are rarely reported in AD, but they both have been associated with the amyloid precursor gene *APP* [49]. *RAD51B* has also been associated with macular degeneration [50,51]. Furthermore, *DISC1* has been associated with late-onset AD [52], as well as Aβ production [53].

4.6 Conclusions

Like in our work here, [43–45] use the ADNI sample. However, these studies use more complex study designs or more complex methods, which in some cases [43,44] demand vast computational resources. Other work, such as [45], also have more data and larger samples. Yet our much simpler and computationally less expensive approach, with a comparable or smaller sample found many of the same effects reported in all of the aforementioned papers, such as the contributions of *GPC6* and *RBFOX1*. Furthermore we also identified many novel effects.

Importantly, PLS-CA found a number well-known genetic contributions to AD such as the effects across Chr19. The Chr19 effects were important because they highlight that our method can and does identify well-known and robust effects, and thus illustrates validity for the problems at hand.

It is clear that PLS-CA has many advantages over traditional and more recent approaches

to GWAS. First, PLS-CA does not make assumptions about genotypic effects; rather, PLS-CA reveals the types of effects (e.g., additive, dominant) and the directions of these effects. For examples our *Validation* phase revealed many complex effects (see also Figure 3 and Table 3): (1) rs769449 is approximately linear where the minor homozygote and the heterozygote are in the risk region, and the major homozygote in the control region, (2) rs4420638 is approximately dominant where the minor allele is in the risk region, (3) rs440666 is approximately recessive and much more importantly, the minor homozygote is in the control region, and (4) rs8052288 showed a heterozygous effect where only the heterozygote was significant and in the risk region. More importantly, because PLS-CA is a multivariate technique it actually provides estimates of the contributions of these genotypes to polygenic effects (see BSRs in Table 3 or component scores in Figure 3). Together these features of PLS-CA are particularly suited for our problem: PLS-CA has helped to identify and specify the expected complex genetic contributions to AD, and thus—hopefully—techniques such as PLS-CA provide the field with an approach much better suited to reduce false positives, non-replications, conflicting reports, and otherwise problematic interpretations of genetics studies in the literature.

4.6.1 Limitations

The ADNI data sets are, by today's standards, relatively small samples for such a study. However our study design emphasized our *Validation* step to help confirm effects identified in *Discovery*. Additionally, our *Validation* phase could have weighted the candidates based on bootstrap ratios from the *Discovery* phase to emphasize the strength of effects. However, we chose not to and opted for a more data-driven strategy. Ultimately this choice was beneficial: for example, if we had used weighted values in the *Validation* phase we may have missed effects

such as those associated with rs6859 (i.e., rs6859 expressed different effects in *Discovery* compared to *Validation*). Finally, the effects we have identified should be further verified in a predictive analysis wherein individuals could be genotyped for the specific markers we have identified; which may be possible through similar studies that are in progress (e.g., ADNI-3, the Ontario Neurodegenerative Disease Research Initiative [54, 55]).

References

[1] Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat Genet 2007;39:168–77. doi:10.1038/ng1943.

[2] Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet 2011;43:436–41. doi:10.1038/ng.801.

[3] Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet 2009;41:1088–93. doi:10.1038/ng.440.

[4] Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet 2009;41:1094–9. doi:10.1038/ng.439.

[5] Bertram L, Tanzi RE. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. Nat Rev Neurosci 2008;9:768–78. doi:10.1038/nrn2494.

[6] Karch CM, Cruchaga C, Goate AM. Alzheimer's Disease Genetics: From the Bench to the Clinic. Neuron 2014;83:11–26. doi:10.1016/j.neuron.2014.05.041.

[7] Schellenberg GD, Montine TJ. The genetics and neuropathology of Alzheimer's disease. Acta Neuropathol (Berl) 2012;124:305–23. doi:10.1007/s00401-012-0996-2.

[8] Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease. Nature 2014;505:550–4. doi:10.1038/nature12825.

[9] van der Lee SJ, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, et al. PLD3 variants in population studies. Nature 2015;520:E2–3. doi:10.1038/nature14038.

[10] Lambert J-C, Grenier-Boley B, Bellenguez C, Pasquier F, Campion D, Dartigues J-F, et al. PLD3 and sporadic Alzheimer's disease risk. Nature 2015;520:E1–E1. doi:10.1038/nature14036.

[11] Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLOS Med 2017;14:e1002258. doi:10.1371/journal.pmed.1002258.

[12] Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Ann Neurol 2017;82:311–4. doi:10.1002/ana.24999.

[13] Hohman TJ, Koran ME, Thornton-Wells T, for the Alzheimer's Neuroimaging Initiative. Epistatic Genetic Effects among Alzheimer's Candidate Genes. PLoS ONE 2013;8:e80839. doi:10.1371/journal.pone.0080839.

[14] Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivières S, Jahanshad N, et al. Common genetic variants influence human subcortical brain structures. Nature 2015;520:224–9. doi:10.1038/nature14101.

[15] Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, et al. Hippocampal Atrophy as a Quantitative Trait in a Genome-Wide Association Study Identifying Novel Susceptibility Genes for Alzheimer's Disease. PLoS ONE 2009;4:e6501. doi:10.1371/journal.pone.0006501.

[16] Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–91. doi:10.1038/nrg1916.

[17] Vormfelde SV, Brockmöller J. On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics. Nat Rev Genet 2007;8. doi:10.1038/nrg1916-c1.

[18] Balding DJ. Reply: On the value of haplotype-based genotype–phenotype analysis and on data transformation in pharmacogenetics and -genomics. Nat Rev Genet 2007;8. doi:10.1038/nrg1916-c2.

[19] Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 2007;31:358–62. doi:10.1002/gepi.20217.

[20] Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J-C, Carrasquillo MM, et al. Common variants in ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat Genet 2011;43:429–35. doi:10.1038/ng.803.

[21] Hooli BV, Lill CM, Mullin K, Qiao D, Lange C, Bertram L, et al. PLD3 gene variants and Alzheimer's disease. Nature 2015;520:E7–8. doi:10.1038/nature14040.

[22] Charney E. Genes, behavior, and behavior genetics. Wiley Interdiscip Rev Cogn Sci 2016:n/a-n/a. doi:10.1002/wcs.1405.

[23] Kauwe JSK, Cruchaga C, Karch CM, Sadler B, Lee M, Mayo K, et al. Fine Mapping of Genetic Variants in BIN1, CLU, CR1 and PICALM for Association with Cerebrospinal Fluid Biomarkers for Alzheimer's Disease. PLOS ONE 2011;6:e15918. doi:10.1371/journal.pone.0015918.

[24] Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, et al. 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 2010;53:1051–63. doi:10.1016/j.neuroimage.2010.01.042.

[25] Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, et al. Voxelwise genome-wide association study (vGWAS). NeuroImage 2010;53:1160–74. doi:10.1016/j.neuroimage.2010.02.032.

[26] Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, et al. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. Neuroimage 2011;56:1875–91. doi:10.1016/j.neuroimage.2011.03.077.

[27] Shulman JM, Chen K, Keenan BT, Chibnik LB, Fleisher A, Thiyyagura P, et al. Genetic susceptibility for Alzheimer disease neuritic plaque pathology. JAMA Neurol 2013;70:1150–1157.

[28] Swaminathan S, Shen L, Risacher SL, Yoder KK, West JD, Kim S, et al. Amyloid pathway-based candidate gene analysis of [11C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Brain Imaging Behav 2011;6:1–15. doi:10.1007/s11682-011-9136-1.

[29] Meda SA, Narayanan B, Liu J, Perrone-Bizzozero NI, Stevens MC, Calhoun VD, et al. A large scale multivariate parallel ICA method reveals novel imaging–genetic relationships for Alzheimer's disease in the ADNI cohort. NeuroImage 2012;60:1608–21. doi:10.1016/j.neuroimage.2011.12.076.

[30] Beaton D, Dunlop J, Abdi H, Alzheimer's Disease Neuroimaging Initiative. Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data. Psychol Methods 2016;21:621–51. doi:10.1037/met0000053.

[31] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

[32] Beaton D, Fatt CRC, Abdi H. An ExPosition of multivariate analysis with the singular value decomposition in R. Comput Stat Data Anal 2014;72:176–89. doi:http://dx.doi.org/10.1016/j.csda.2013.11.006.

[33] Beaton D, Rieck J, Abdi H. TInPosition: Inference Tests for TExPosition. 2013.

[34] Abdi H. Discriminant correspondence analysis. Encycl. Meas. Stat., 2007, p. 284–90.

[35] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007;81:559–75.

[36] Melville S. NCBI2R: Queries and Annotates SNPs, Genes and Pathway from NCBI. 2015.

[37] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 2009;4:1184–91. doi:10.1038/nprot.2009.97.

[38] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinforma Oxf Engl 2005;21:3439–40. doi:10.1093/bioinformatics/bti525.

[39] Efron B. Bootstrap Methods: Another Look at the Jackknife. Ann Stat 1979;7:1–26.

[40] Hao X, Li C, Du L, Yao X, Yan J, Risacher SL, et al. Mining Outcome-relevant Brain Imaging Genetic Associations via Three-way Sparse Canonical Correlation Analysis in Alzheimer's Disease. Sci Rep 2017;7:srep44272. doi:10.1038/srep44272.

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

[42] Zhao J, Zhu Y, Yang J, Li L, Wu H, De Jager PL, et al. A genome-wide profiling of brain DNA hydroxymethylation in Alzheimer's disease. Alzheimers Dement 2017;13:674–88. doi:10.1016/j.jalz.2016.10.004.

[43] Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. NeuroImage 2012;60:700–16. doi:10.1016/j.neuroimage.2011.12.029.

[44] Huang C, Thompson P, Wang Y, Yu Y, Zhang J, Kong D, et al. FGWAS: Functional genome wide association analysis. NeuroImage 2017;159:107–21. doi:10.1016/j.neuroimage.2017.07.030.

[45] Sherva R, Tripodis Y, Bennett DA, Chibnik LB, Crane PK, de Jager PL, et al. Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. Alzheimers Dement 2014;10:45–52. doi:10.1016/j.jalz.2013.01.008.

[46] Gillingwater TH, Wishart TM. Mechanisms underlying synaptic vulnerability and degeneration in neurodegenerative disease. Neuropathol Appl Neurobiol 2013;39:320–34. doi:10.1111/nan.12014.

[47] Zhang M, Mu H, Shang Z, Kang K, Lv H, Duan L, et al. Genome-wide pathway-based association analysis identifies risk pathways associated with Parkinson's disease. Neuroscience 2017;340:398–410. doi:10.1016/j.neuroscience.2016.11.004.

[48] Durrenberger PF, Fernando FS, Kashefi SN, Bonnert TP, Seilhean D, Nait-Oumesmar B, et al. Common mechanisms in neurodegeneration and neuroinflammation: a BrainNet Europe gene expression microarray study. J Neural Transm 2015;122:1055–68. doi:10.1007/s00702-014-1293-0.

[49] Yokota T, Mishra M, Akatsu H, Tani Y, Miyauchi T, Yamamoto T, et al. Brain site-specific gene expression analysis in Alzheimer's disease patients. Eur J Clin Invest 2006;36:820–30. doi:10.1111/j.1365-2362.2006.01722.x.

[50] Consortium TAG. Seven new loci associated with age-related macular degeneration. Nat Genet 2013;45:433–9. doi:10.1038/ng.2578.

[51] Logue MW, Schu M, Vardarajan BN, Farrell J, Lunetta KL, Jun G, et al. A search for age-related macular degeneration risk variants in Alzheimer disease genes and pathways. Neurobiol Aging 2014;35:1510.e7-1510.e18. doi:10.1016/j.neurobiolaging.2013.12.007.

[52] Beecham GW, Martin ER, Li Y-J, Slifer MA, Gilbert JR, Haines JL, et al. Genome-wide Association Study Implicates a Chromosome 12 Risk Locus for Late-Onset Alzheimer Disease. Am J Hum Genet 2009;84:35–43. doi:10.1016/j.ajhg.2008.12.008.

[53] Shahani N, Seshadri S, Jaaro-Peled H, Ishizuka K, Hirota-Tsuyada Y, Wang Q, et al. DISC1 regulates trafficking and processing of APP and Aβ generation. Mol Psychiatry 2014. doi:10.1038/mp.2014.100.

[54] Farhan SMK, Bartha R, Black SE, Corbett D, Finger E, Freedman M, et al. The Ontario Neurodegenerative Disease Research Initiative (ONDRI). Canadian Journal of Neurological Sciences 2016:1–7. doi:10.1017/cjn.2016.415.

[55] Farhan SMK, Dilliott AA, Ghani M, Sato C, Liang E, Zhang M, et al. The ONDRISeq panel: custom-designed next-generation sequencing of genes related to neurodegeneration. Npj Genomic Medicine 2016;1:16032. doi:10.1038/npjgenmed.2016.32.

Acknowledgements

Data collection and sharing for this project was funded by the 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 the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; 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.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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. For up-to-date information and data see: http://www.adni-info.org and http://adni.loni.usc.edu/

Supplemental Material

Table S1
Dx GWAS Significant genotypes

(a) Dx GWAS Component 1 (AD vs. MCI) – significant genotypes

rsid	genotype	BSR –1	BSR –2	Chr	Gene Symbol
rs1061622	GG	-5.320	-1.059	1	TNFRSF1B
rs6696924	GG	-5.516	-0.177	1	C8A
rs679350	TT	-5.007	-0.244	1	C8A
rs1522551	TT	-5.075	-0.197	3	
rs7626449	AA	-5.332	0.454	3	
rs13157174	GG	-5.170	-0.074	5	LINC02216
rs2875382	TT	-5.057	-1.290	6	MRAP2
rs4501410	GG	-6.050	1.185	6	
rs11783013	GG	-5.498	-0.397	8	
rs902466	TT	-5.440	0.230	10	ARHGAP19-SLIT1
rs16977252	GG	-5.617	-0.471	15	AKAP13
rs12934725	GG	-5.230	0.536	16	RBFOX1
rs9952815	GG	-6.027	0.745	18	NOL4
rs283168	TT	-5.331	-0.389	19	
rs916326	GG	5.356	0.315	20	PTPRT
rs13054435	AA	-5.541	-0.213	22	NUP50-AS1
rs4820946	CC	-5.641	-0.500	22	MIR3928

(b) Dx GWAS Component 2 (CON vs. {AD & MCI}) – significant genotypes

rsid	genotype	BSR - 1	BSR - 2	Chr	Gene
rs1115763	CC	-1.001	5.970	2	AC007682.1
rs12618595	AC+AA	1.418	5.474	2	OSBPL6
rs1470524	TT	-1.144	5.178	2	
rs1387089	CT+CC	-0.701	5.015	3	
rs1387094	GT+GG	-0.541	5.004	3	
rs4685465	GT+GG	-0.765	5.136	3	
rs9821034	GG	0.905	5.370	3	
rs3846336	AG+AA	0.462	5.329	4	CCDC149
rs6448119	TT	0.027	5.711	4	
rs6812046	TT	0.355	5.462	4	
rs6851636	GG	-0.371	5.457	4	
rs7656406	AA	0.184	5.480	4	
rs6882277	GG	0.230	5.867	5	
rs13236754	GG	-0.163	5.234	7	ZNF398
rs1557664	GG	1.430	5.924	7	
rs9640538	AA	-0.297	5.487	7	
rs10814567	GG	0.160	5.360	9	POLR1E
rs10814571	AA	-0.128	6.816	9	POLR1E
rs1590255	CA+CC	-0.732	5.042	9	
rs3739574	TT	-0.121	6.805	9	POLR1E
rs7468695	GA+GG	-1.168	5.073	9	
rs11598825	AA	-0.487	5.257	10	ITIH5
rs947696	TT	1.400	5.473	10	
rs11160481	CC	0.013	5.017	14	
rs1243473	TT	0.439	5.782	14	ARHGEF40
rs4635275	AA	0.002	5.028	14	
rs7146951	GG	0.005	5.764	14	
rs9806693	AA	-0.668	5.969	15	MORF4L1
rs4789240	TT	1.051	5.089	17	SDK2
rs9892996	AA	-1.531	6.928	17	
rs9949152	CC	-3.019	5.005	18	
rs2075650	AA	3.295	-5.584	19	TOMM40
rs7251241	AA	0.937	5.282	19	UNC13A
rs2830052	CC	-0.086	5.614	21	APP
rs8141950	TC+TT	0.767	5.674	22	PARVB

Note. Significant genotypes from Dx GWAS Components 1 (a) and 2 (b). Gene symbols via `NCBI2R`. Component 1 separated AD from MCI, thus, most genotypes here are more associated

with AD than MCI; the exception is rs916326 where GG is more associated with MCI than AD. Component 2 separates CON from {AD & MCI}, thus, most genotypes here are more associated with disorder status (AD or MCI) than CON; the exception is rs2075650 where AA is more associated with CON than disease status.

Table S2
ApoE GWAS Significant genotypes

(a) ApoE GWAS Component 1 (Presence vs. Absence of E4) – significant genotypes

rsid	genotype	BSR - 1	BSR - 2	Chr	Gene Symbol
rs10830213	GG	6.537	-0.724	11	RAB38
rs157580	AA	-7.842	-2.918	19	TOMM40
rs157580	GG	14.612	-2.130	19	TOMM40
rs2075650	AA	15.323	-0.619	19	TOMM40
rs405509	AA	-6.101	-2.567	19	APOE
rs405509	CC	6.701	-1.244	19	APOE
rs439401	CC	-8.703	-2.630	19	
rs6859	AA	-5.614	-3.096	19	NECTIN2/PVRL2
rs8106922	AA	-7.514	-3.490	19	TOMM40
rs8106922	GG	20.928	-4.430	19	TOMM40

(b) ApoE GWAS Component 2 (2 E4 alleles vs. {0 or 1}) – significant genotypes

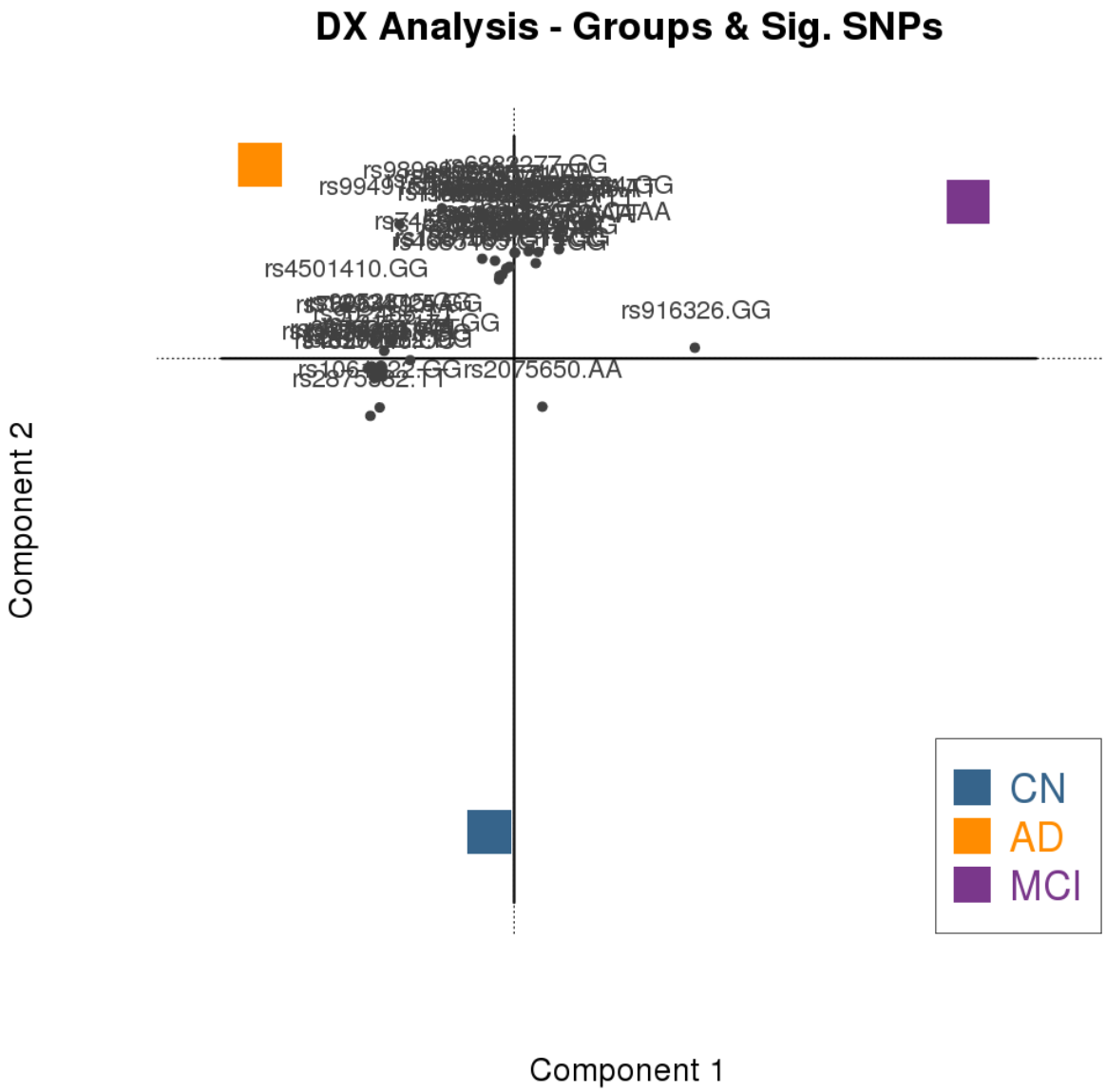
rsid	genotype	BSR - 1	BSR - 2	Chr	Gene Symbol
rs2000072	AA	-0.211	5.066	1	LINC00624
rs4253890	CC	-1.077	5.985	1	PTPN14
rs6681032	TC+TT	0.157	5.439	1	
rs6703696	GA+GG	0.169	5.604	1	
rs7541019	GG	0.576	5.069	1	TSNAX-DISC1,DISC1
rs11899115	AA	-0.640	5.996	2	
rs13009482	CT+CC	-1.383	5.696	2	
rs722963	TT	-0.288	5.093	2	
rs10019637	CC	-0.964	5.412	4	
rs10804966	AA	-0.151	5.267	4	EVC
rs300574	TT	0.275	5.860	4	SPRY1
rs7678082	TC+TT	-0.105	5.196	4	WWC2
rs7681283	GG	-0.168	5.278	4	EVC
rs868082	TC+TT	-0.187	6.130	4	
rs10041935	CC	-0.895	5.750	5	
rs236444	TT	-1.938	7.117	6	CPNE5
rs1673206	TT	-1.294	5.026	7	
rs6960851	AA	-0.713	5.079	7	
rs537941	TT	-0.730	5.270	8	NCALD
rs1492598	AA	-0.933	5.119	9	
rs4935847	TT	-0.247	6.085	11	
rs1647147	GG	-1.396	5.407	12	
rs16928445	TT	0.609	5.015	12	LRMP
rs4759955	TT	-0.490	5.903	12	TMEM132D
rs6582412	AA	-1.433	5.113	12	
rs944838	CC	0.373	5.917	13	GPC6
rs9549831	AA	0.454	5.082	13	
rs4905290	GG	-0.457	5.019	14	CLMN
rs6573852	GG	0.438	5.294	14	RAD51B
rs7148010	TT	1.018	5.112	14	SMOC1
rs10519492	GA+GG	-0.468	5.537	15	
rs6496431	GG	-0.087	5.072	15	
rs714900	TC+TT	-0.627	5.611	15	
rs7177541	AA	-0.664	5.299	15	
rs4500815	AA	-0.280	5.300	18	CTIF
rs9955327	CC	-1.958	5.021	18	CELF4
rs10423685	TT	-0.479	5.053	19	ZNF600
rs157580	GA	4.603	8.591	19	TOMM40
rs6509238	CC	0.056	5.038	19	

rs8106922	GA	3.703	9.456	19	TOMM40

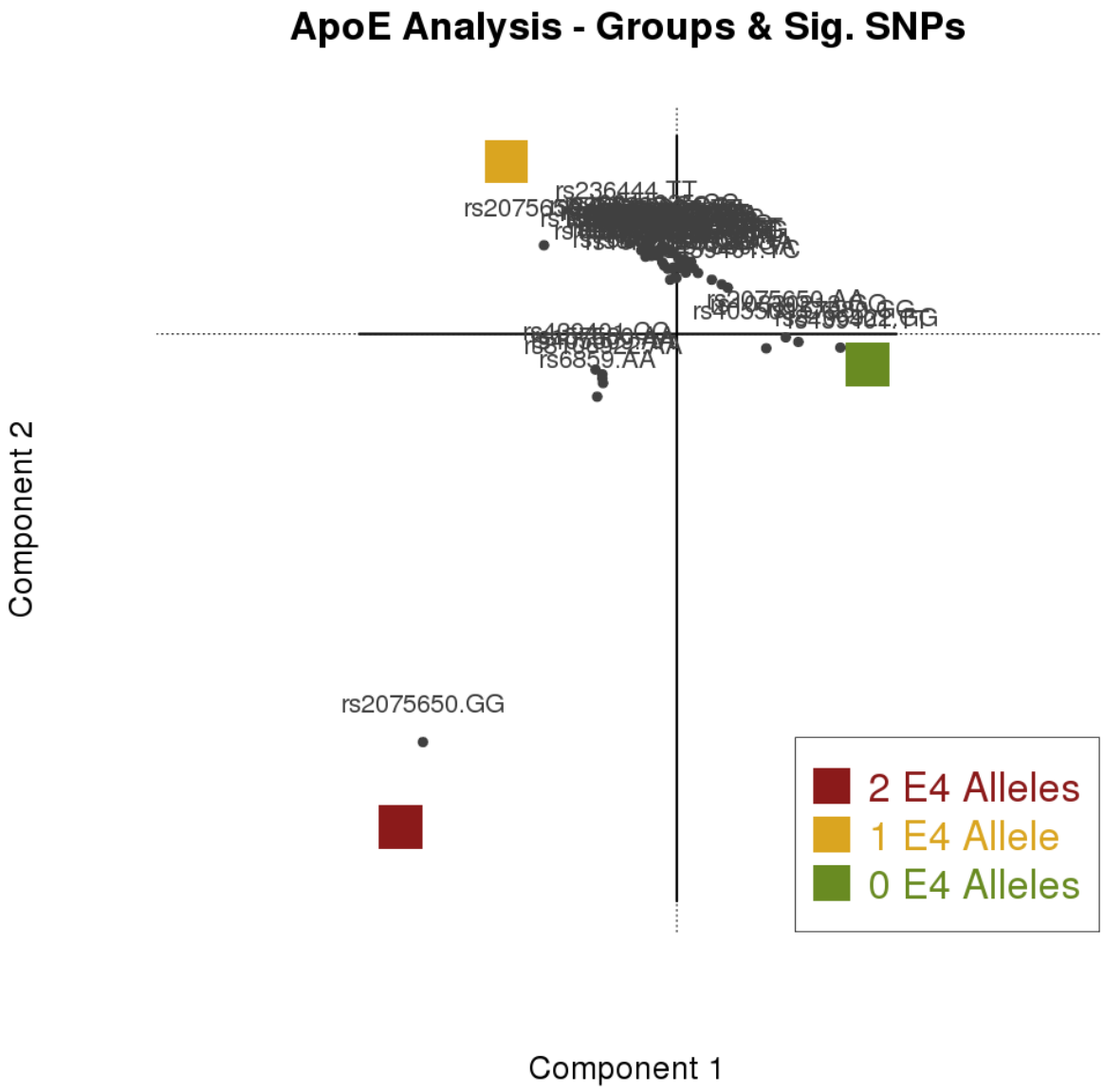
(c) ApoE GWAS Significant genotypes on both Components.

rsid	genotype	BSR - 1	BSR - 2	Chr
rs2075650	GA	-17.593	7.374	19
rs2075650	GG	-25.132	-13.27	19
rs439401	TC	5.052	7.920	19
rs439401	TT	25.848	-5.790	19

Note. Significant genotypes from ApoE GWAS Components 1 (a), 2 (b), and on both Components (c). Gene symbols via `NCBI2R`. Component 1 separated presence from absence of E4 alleles. Genotypes that have a negative bootstrap ratio (BSR) were more associated with the presence of an E4 allele. Component 2 separated, essentially, the 2 E4 alleles from the other (0 or 1) E4 alleles. In (b) all genotypes were more related to the *absence* of 2 E4 alleles. In (c) these genotypes contribute to both components and suggest that these genotypes are in very high linkage disequilibrium with ApoE (note that the GG genotype of rs2075650 strongly contributes to both components and in the same direction as the *presence* 2 E4 alleles).



Supplemental Figure 1. The component map from the Dx-GWAS (discovery phase). The component map shows significant genotypes and the group configuration to illustrate—as a biplot—the relationship between the genotypes and groups.



Supplemental Figure 2. The component map from the ApoE-GWAS (discovery phase). The component map shows significant genotypes and the group configuration to illustrate—as a biplot—the relationship between the genotypes and groups.