bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Whole Exome Sequencing in 20,197 Persons for Rare Variants in Alzheimer Disease

Neha S. Raghavan, PhD [1,2], Adam M. Brickman, PhD[1,2,3], Howard Andrews, PhD[1,2,4], Jennifer J. Manly, PhD[1,2,3], Nicole Schupf, PhD[1,2,3], Rafael Lantigua, MD[1,6], The Alzheimer's Disease Sequencing Project*, Charles J. Wolock, BA[8], Sitharthan Kamalakaran, PhD[8], Slave Petrovski, PhD[8,9], Giuseppe Tosto, MD, PhD [1,2], Badri N. Vardarajan, PhD[1,2,3], David B. Goldstein, PhD[3,6,8] and Richard Mayeux, MD[1,2,3,4,7]

KEYWORDS: Alzheimer's disease, Mutations, Whole Exome Sequencing

RUNNING TITLE: Ultra-Rare Variants and Alzheimer's disease

WORD COUNT: Abstract 250; Introduction 344; Discussion 1,125; Body 2,440

FIGURES AND TABLES: Figures 1; Tables 4; Supplemental Tables 2

CHARACTERS: Title 122; Running Head 43

[1]The Taub Institute for Research on Alzheimer's Disease and the Aging Brain, [2]The Gertrude H. Sergievsky Center, The Departments of [3]Neurology, [4]Psychiatry, [5]Systems Biology and [6]Medicine, at the College of Physicians and Surgeons, Columbia University, The New York Presbyterian Hospital, [7]The Department of Epidemiology, Mailman School of Public Health, [8]Institute of Genomic Medicine, Columbia University, The New York Presbyterian Hospital, New York, NY, USA, [9]AstraZeneca Centre for Genomics Research, Precision Medicine and Genomics, IMED Biotech Unit, AstraZeneca, Cambridge CB2 0AA, UK

*Members listed by Institution in the supplement

Correspondence:
Richard Mayeux, MD, MSc.
Department of Neurology
710 West 168th Street
Columbia University
New York, NY 10032
Phone: 212-305-2391
Email: rpm2@cumc.columbia.edu

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

1 **Abstract**

2 **Objective**

3 The genetic bases of Alzheimer's disease remain uncertain. An international effort to fully
4 articulate genetic risks and protective factors is underway with the hope of identifying potential
5 therapeutic targets and preventive strategies. The goal here was to identify and characterize
6 the frequency and impact of rare and ultra-rare variants in Alzheimer's disease using whole
7 exome sequencing in 20,197 individuals.
8

9 **Methods**

10 We used a gene-based collapsing analysis of loss-of-function ultra-rare variants in a case-
11 control study design with data from the Washington Heights-Inwood Columbia Aging Project,
12 the Alzheimer's Disease Sequencing Project and unrelated individuals from the Institute of
13 Genomic Medicine at Columbia University.
14

15 **Results**

16 We identified 19 cases carrying extremely rare *SORL1* loss-of-function variants among a
17 collection of 6,965 cases and a single loss-of-function variant among 13,252 controls ($p = 2.17 \times$
18 10^{-8}; OR 36.2 [95%CI 5.8 - 1493.0]). Age-at-onset was seven years earlier for patients with
19 *SORL1* qualifying variant compared with non-carriers. No other gene attained a study-wide
20 level of statistical significance, but multiple top-ranked genes, including *GRID2IP*, *WDR76* and
21 *GRN,* were among candidates for follow-up studies.
22

23 **Interpretation**

24 This study implicates ultra-rare, loss-of-function variants in *SORL1* as a significant genetic risk
25 factor for Alzheimer's disease and provides a comprehensive dataset comparing the burden of
26 rare variation in nearly all human genes in Alzheimer's disease cases and controls. This is the

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

27 first investigation to establish a genome-wide statistically significant association between
28 multiple extremely rare loss-of-function variants in *SORL1* and Alzheimer's disease in a large
29 whole-exome study of unrelated cases and controls.
30

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

31 **Introduction**

32 Alzheimer's disease (AD) is a highly prevalent disorder that dramatically increases in frequency
33 with age, and has no effective treatment or means of prevention. While three causal genes,
34 Amyloid Precursor Protein (*APP*), Presenilin 1 and 2 (*PSEN1* and *PSEN2*), have been
35 established for early-onset AD (age of onset <65 years of age), the rest of the heritability is still
36 unknown. Further, beyond Apolipoprotein E (*APOE*), which confers the greatest risk for late-
37 onset AD (age of onset ≥65 years of age), there remains a large gap in the understanding of its
38 causes. Identifying genetic variants that increase risk or protect against AD is considered an
39 international imperative because of the potential therapeutic targets that may be revealed.
40 Recent technological advances in genome-wide association studies and high throughput next-
41 generation sequencing may help to implicate variants in genes in specific molecular pathways
42 relevant to AD.
43
44 In this study, we used whole-exome sequencing to investigate all protein-coding genes in the
45 genome focusing on ultra-rare (allele frequency less than 0.01%) and putatively deleterious
46 variants. Rare variants are hypothesized to contribute to disease [1, 2], and studies of complex
47 traits in population genetic models indicate an inverse relationship between the odds ratio and
48 effect size conferred by rare variants and low allele frequencies [3]. Thus, we searched for large
49 effects conferred by putatively causal ultra-rare variants. Traditional single variant statistics can
50 be underpowered because patients with similar clinical presentations possess distinct rare
51 variants that inflict similar effects on the gene [4]. Gene-based collapsing analyses increase signal
52 detection by aggregating individual qualifying variants within an *a priori* region (e.g., a gene),
53 facilitating detection of genes associated with disease through a specific class of genetic
54 variation (e.g., loss-of-function variants).
55

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

56 In order to maximize the ability to detect ultra-rare variants associated with AD, exome-
57 sequencing data of 20,197 cases and controls from the Washington Heights-Inwood Community
58 Aging Project (WHICAP), the Alzheimer's Disease Sequencing Project (ADSP) and unrelated
59 controls from the Institute of Genomic Medicine were systematically combined and analyzed
60 using a collapsing method with proven prior success in identifying disease associated genes [5, 6].
61

62 Methods

63 The three groups used in this study and their sequencing information are described below.
64 **Washington Heights-Inwood Community Aging Project.** The WHICAP study consisted of a
65 multi-ethnic cohort of 4,100 individuals followed over several years The cohort participants were
66 non-demented initially, 65 years of age or older, and comprised of non-Hispanic whites, African
67 Americans, and Caribbean Hispanics from the Dominican Republic. During each assessment,
68 participants received a neuropsychological test battery, medical interview, and were re-
69 consented for sharing of genetic information and autopsy. A consensus diagnosis was derived
70 for each participant by experienced clinicians based on NINCDS-ADRDA criteria for possible,
71 probable, or definite AD, or moderate or high likelihood of neuropathological criteria of AD [7,
72 8]. Every individual with whole-exome sequencing has at least a baseline and one follow-up
73 assessment and examination, and for those who have died, the presence or absence of
74 dementia was determined using a brief, validated telephone interview with participant
75 informants: the Dementia Questionnaire (DQ) [9] and the Telephone Interview of Cognitive Status
76 (TICS) [10]. 3,702 exome-sequenced WHICAP individuals were designated with case or control
77 status and included in this analysis. From the sequenced cohort, 27% died and less than 1%
78 were lost at follow-up.
79

80 **Alzheimer's Disease Sequencing Project.** The ADSP, developed by the National Institute on
81 Aging (NIA) and National Human Genome Research Institute (NHGRI) includes a large case-

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

82 control cohort of approximately 10,000 individuals [7]. The recruitment of these individuals was in
83 collaboration with the Alzheimer's Disease Genetics Consortium and the Cohorts for Heart and
84 Aging Research in Genomic Epidemiology Consortium. The details and rationale for the case-
85 control selection process have been previously described [7]. All cases and controls were at least
86 60 years old and were chosen based on sex, age and *APOE* status: 1) controls were evaluated
87 for their underlying risk for AD and for their likelihood of conversion to AD by age 85, based on
88 age at last examination, sex, and *APOE* genotype, and those with the least risk for conversion
89 to AD were selected, and 2) cases were evaluated for their underlying risk for AD based on age
90 at onset, sex, and *APOE* genotype and those with a diagnosis least explained by these factors
91 were selected [7]. Cases were determined either because they met NINCDS-ADRDA clinical
92 criteria for AD, or postmortem findings met moderate or high likelihood of neuropathological
93 criteria of AD [7, 8]. Autopsy data was available for 28.7% of the cases and controls used in the
94 analysis. Further, some cases were originally diagnosed clinically, subsequently died and had
95 neuropathological findings available after postmortem examination. Cases had documented age
96 at onset or age at death (for pathologically determined cases). Controls were free of dementia
97 by direct, documented cognitive assessment or neuropathological results. The ADSP group
98 consisted of European-Americans and Caribbean Hispanics. All data were available for
99 download for approved investigators at The National Institute on Aging Genetics of Alzheimer's
100 Disease Data Storage Site website (https://www.niagads.org/adsp/content/home). As part of the
101 ADSP, 116 non-Hispanic white WHICAP controls and 34 cases previously sequenced were
102 included here.
103
104 **Additional Controls.** The Institute for Genomic Medicine (IGM) (Columbia University Medical
105 Center, New York, NY) hosts an internal database of sequencing data collected from previously
106 exome-sequenced material. In this study, exome-sequencing data from 6,395 IGM controls
107 were utilized. All data used were previously consented for future control use from multiple

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

108 studies of various phenotypes. The cohort was made up of 55.7% healthy controls and 46.3%
109 with diseases not co-morbid with AD (disease classifications shown in Supplemental Table 1).
110 Although the cohort of controls were not enriched for any neurological disorder or diseases with
111 a known co-morbidity with AD, presence or future possibility of AD could not be excluded based
112 on the available clinical data. individuals with Age and *APOE* status were not available for these
113 participants. The cohort comprised of 70% non-Hispanic white individuals along with those of
114 African American, Hispanic, Middle Eastern, Asian and unknown descent.
115
116 **Sequencing, Quality Control and Variant Calling**
117 Whole-exome sequencing of the WHICAP cohort was performed at Columbia University. The
118 additional controls were sequenced at Duke University and Columbia University. Whole-exome
119 sequencing of the ADSP cohort was performed at The Human Genome Sequencing Center,
120 Baylor College of Medicine, Houston, Texas; The Broad Institute Sequencing Platform, The Eli
121 & Edythe L. Broad Institute of the Massachusetts Institute of Technology and Harvard
122 University, Cambridge Massachusetts and Washington University Genome Sequencing Center,
123 Washington University School of Medicine, Saint Louis, Missouri. ADSP raw files in the
124 sequencing read archive format were downloaded from the dbGAP database and
125 decompressed to obtain FASTQ files.
126
127 All data were reprocessed for a consistent alignment and variant calling pipeline consisting of
128 the primary alignment and duplicate marking using the Dynamic Read Analysis for Genomics
129 (DRAGEN) platform followed by variant calling according to best practices outlined in Genome
130 Analysis Tool Kit (GATK v3.6). Briefly, aligned reads were processed for indel realignment
131 followed by base quality recalibration and Haplotype calling to generate variant calls. Variant
132 calls were then subject to Variant Quality Score Recalibratrion (VQSR) using the known single
133 nucleotide variants (SNVs) sites from HapMap v3.3, dbSNP, and the Omni chip array from the

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

134 1000 Genomes Project. SNVs were required to achieve a tranche of 99.9% and indels a
135 tranche of 95%. Finally, read-backed phasing was performed to determine phased SNVs and
136 merge multinucleotide variants (MNVs) when appropriate. Variants were annotated using Clin-
137 Eff with Ensembl-GRCh37.73 annotations.
138
139 Quality thresholds were set based on previous work[5, 6], such that all resulting exome variants
140 had a quality score of at least 50, quality by depth score of at least 2, genotype quality score of
141 at least 20, read position rank sum of at least -3, mapping quality score of at least 40, mapping
142 quality rank sum greater than -10, and a minimum coverage of at least 10. SNVs had a
143 maximum Fisher's strand bias of 60, while indels had a maximum of 200. For heterozygous
144 genotypes, the alternative allele ratio was required to be greater than or equal to 25% and
145 variant from sequencing artifacts and exome variant server failures
146 (http://evs.gs.washington.edu/EVS) were excluded.
147
148 Quality control was performed on all sequencing data. Samples with less than 90% of the
149 consensus coding sequence (CCDS) covered at 10X and samples with sex-discordance
150 between clinical and genetic data were excluded from the analysis. Cryptic relatedness testing
151 was performed using KING, and second degree or closer (relatedness threshold of 0.0884 or
152 greater) relatives were removed with preferential retention of cases over controls and
153 subsequently samples with higher average read-depth coverage.
154
155 The consensus coding sequence [11] (CCDS) annotated protein-coding region for each gene
156 (n=18,834) was tabulated as either carrying or not carrying a qualifying variant for every
157 individual. Qualifying variants were defined for a loss-of-function model: stop gain, frameshift,
158 splice site acceptor, splice site donor, start lost, or exon deleted variants. A negative control
159 analysis was performed defining qualifying variants as synonymous variants to detect potential

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

160 biases in variant calling between the cases and controls separately for each of the top four
161 genes. The minor allele frequency threshold was 0.01% internally and within African American,
162 Latino and Non-Finnish European populations from the Exome Aggregation Consortium[12]
163 (ExAC release version 0.3.1). The allele frequency thresholds use a "leave-one-out" method for
164 the combined test cohort of cases and controls such that the minor allele frequency of each
165 variant was calculated using all individuals except for the index sample under investigation.
166 Thus, the maximum instances of a single variant a gene in our sample of 20,197 was five. A
167 dominant model was defined such that one or more qualifying variant(s) in a gene qualified the
168 gene.

169

170 An important aspect of the collapsing analysis methodology is the reduction of variant calling
171 bias due to coverage differences between cases and controls. To ensure balanced sequencing
172 coverage of evaluated sites between cases and controls, we imposed a statistical test of
173 independence between the case/control status and coverage. For a given site, consider s total
174 number of cases, t total number of controls and x number of cases covered at 10X, y number of
175 controls covered at 10x. We model the number of covered cases X as a Binomial random
176 variable:

177 $$X \sim \text{bin}(n = \text{number covered samples}, p = P(\text{case}|\text{covered}))$$

178 If case/control status and coverage status are independent, then:

179 $$P\ (\text{case}|\text{covered}) = P\ (\text{case}) = s/(s+t)$$

180 We can test for this independence by performing a two-sided Binomial test on the number of
181 covered samples at given site, x.

182 $$\text{BinomTest}(k=x,\ n=x+y,\ p=s/(\ s+t))$$

183

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

184 In the collapsing analyses, a binomial test for coverage balance as described above was
185 completed as an additional qualifying criterion. Any site which resulted in a nominal significance
186 threshold of 0.05 was eliminated from further consideration.
187 A Fisher's exact test on qualifying variants in cases and controls for each gene was performed
188 and imbalances in cases and controls within a gene indicated a possible association with the
189 case-ascertained phenotype. Ultra-rare variant analyses were conducted using Analysis Tools
190 for Annotated Variants (ATAV), developed and maintained by the Institute for Genomic
191 Medicine at Columbia University. Study-wise significance was set to 0.05/18,834(# of genes
192 tested) = $2.7x10^{-6}$. Fisher's Exact Test for the polygenic comparison of International Genetics of
193 Alzheimer's Project (IGAP) loci [13] and t-test for age of onset-analysis (presented as mean +/-
194 standard deviation) were conducted in R v.3.3.1.
195

196 **Results**

197 We analyzed the exomes of 6,965 individuals meeting with the diagnosis of AD and 13,232
198 controls **(Table 1)**. Prior to analysis, 570 individuals (91 cases and 479 controls) were removed
199 due to known or cryptic relatedness. For ultra-rare variant analysis (MAF of 0.01% or lower),
200 conventional population stratification has not been a strong confounder as it can be in common
201 variant analyses; and these results did not significantly differ from meta-analyses in population
202 stratified data. All variants reported here were found in five or less individuals from the study,
203 and most variants were found in only one person, increasing the confidence that population
204 stratification was not an issue. An important distinction exists between the cases and controls in
205 the ADSP and WHICAP datasets. In the ADSP dataset, the younger cases were preferentially
206 chosen as part of the study design [7]. The WHICAP individuals are part of a population-based
207 cohort followed longitudinally, and thus cases were older than controls.
208

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

209 Of the 18,834 genes analyzed, 15,736 contained at least one qualifying variant. Genomic
210 inflation for the analysis was very modest, $\lambda = 1.04$ (**Figure 1**). Gene-based, collapsing analyses
211 for loss-of-function variants, with allele frequency less than 0.01% (within the study cohort, and
212 separately within ExAC[12]) identified *SORL1* to be enriched in cases compared to controls at an
213 exome-wide significance level of $p = 2.17 \times 10^{-8}$ (**Table 2**). We confirmed the results for *SORL1*
214 were not driven by a particular ethnicity by running individual association tests on non-Hispanic
215 Whites, Caribbean Hispanics, and African Americans as described above, separately and
216 summarizing them in a sample weight meta-analysis[14] (*SORL1* $p = 2.45 \times 10^{-8}$). Although no
217 other gene attained the study-wide level of statistical significance, *GRID2IP* ($p = 2.98 \times 10^{-4}$),
218 *WDR76* ($p = 7.39 \times 10^{-4}$) and *GRN* ($p = 9.56 \times 10^{-4}$) were highly-ranked candidate genes that
219 were case-enriched for loss-of-function variants (**Table 2**). Extended results are found in
220 **Supplemental Table 2**. There were no significant differences in synonymous variation in these
221 four genes (1.5% cases, 1.7% of controls; FET $p = 0.25$).
222
223 There were 19 cases with a loss-of-function qualifying variant in *SORL1* (**Table 3**) among 6,965
224 cases (frequency = 0.27%) and one variant among 13,232 controls (frequency = 0.0076%).
225 Given the rate of *SORL1* loss-of-function qualifying variants found in our control sample (1 /
226 13,232; frequency = 0.0076%), we expected to identify only 0.5 loss-of-function variants by
227 chance among our 6,965 cases; however, we identified 19. The accompanying odds ratio for
228 AD risk upon identifying a *SORL1* loss-of-function qualifying variants as defined in this study
229 was 36 [95% CI 5.8 – 1493.0]. Targeted investigation into the single control indicated a
230 diagnosis of mild cognitive impairment[15]. The *SORL1* loss-of-function variants were found
231 across the non-Hispanic white, Caribbean Hispanic, and African American cases. Six of the 19
232 cases were deceased with autopsy confirmation of the AD diagnosis[16].
233

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

234 Of relevance to loss-of-function variant case-enrichment, *SORL1* is known to be among the
235 protein-coding genes most significantly depleted of loss-of-function variants in the general
236 population (LOF depletion FDR = $2x10^{-7}$) (**Table 2**). Of the 17 distinct *SORL1* loss-of-function
237 qualifying variants, only one (11:121440980, rs200504189) was found in the ExAC database[12].
238 *SORL1* was also significantly enriched for functional variants (nonsynonymous and predicted as
239 possibly or probably damaging by PolyPhen-2 HumVar[17]) ($p = 9.79x10^{-7}$), 1.8% of cases had a
240 qualifying functional variant compared to 1% controls. There was no difference in the frequency
241 of *APOE-ε4* carriers among cases with qualifying variants in *SORL1* compared to those without
242 these variants (40.0% vs. 39.6%). Age-at-onset analyses revealed a 6.81 year difference
243 between cases with a *SORL1* qualifying variant versus non-carrying cases (AD carriers: 69.86
244 +/- 9.37; AD non-carriers: 76.67 +/- 8.53; t(6963), $p = 4x10^{-4}$).
245
246 Coverage for the 12 qualifying *GRID2IP* variants was lower in the sequencing performed in this
247 project and in ExAC[12], reducing our confidence of the rare variant calling for this gene because
248 it is likely not represented well by exome capture libraries. The median of mean read-depth
249 coverage of the *GRID2IP* variants was 21-fold and at these exact same sites in ExAC[12], 4-fold.
250 However, read-depth coverage was higher in the genome aggregation database (gnomAD),
251 with a median of mean read-depth coverage of 21-fold, and only two loss-of-function variants
252 less than the 0.0001 allele frequency threshold. Two of the 11 cases were deceased with
253 autopsy confirming the pathological diagnosis of AD[16].
254
255 Coverage for *WDR76* and *GRN* were excellent in this study and in ExAC[12]. Three of the 10
256 individuals clinically diagnosed as AD with loss-of-function qualifying variants in *WDR76* had
257 undergone autopsy. One met postmortem criteria defined as high likelihood of Alzheimer's
258 disease, a second met intermediate likelihood[16], however, the third had no distinctive pathology
259 and no definitive diagnosis was derived. Two of the 11 individuals with *GRN* loss-of-function

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

260 qualifying variants had autopsy data; one met criteria for AD and the other for frontotemporal
261 lobar degeneration (FTLD) [18]. None of the GRN carriers carried variants in any of the top four
262 genes.

263 We also investigated rare variants in loci that were associated with AD in the IGAP genome
264 wide association study [13] along with *APP*, *PSEN1*, *PSEN2*, and *TREM2*. (**Table 4).** Qualifying
265 variants in *SORL1* and *ZCWPW1* (p=0.02) were more frequent in cases than controls. Overall,
266 there was a slight increase in the frequency of variants in cases compared with controls
267 (Fisher's exact p=0.002), but after the removal of *SORL1*, the association was no longer
268 significant (Fisher's exact p=0.11).

269 **Discussion**

270 This study provides strong evidence that ultra-rare, loss-of-function variants in *SORL1* represent
271 an important genetic risk factor for AD. This is the first investigation to establish a genome-wide
272 statistically significant association between ultra-rare variants in *SORL1* and AD in a large,
273 unbiased whole-exome study of unrelated early- and late-onset cases and controls. *SORL1* has
274 previously been implicated in both familial and sporadic, early- and late-onset Alzheimer's
275 disease [19-25].

276 Common variants in *SORL1* were first genetically associated with AD in a candidate gene
277 analysis using 29 common variants [24]. Shortly thereafter, nine rare loss-of-function variants
278 including nonsense, frameshift and splice site mutations were described in familial and sporadic
279 early onset AD [19, 20]. The *SORL1* findings in early onset AD were replicated in larger European
280 cohorts of patients[21]. Using a targeted, candidate gene approach, *SORL1* variants were found
281 by us in familial and sporadic late-onset AD among Caribbean Hispanics as well as patients with
282 European ancestry with sporadic late-onset AD [26]. Our findings here indicated that cases who
283 possess a *SORL1* qualifying variant were on average younger at onset. Yet, only four of the

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

284 cases with a *SORL1* qualifying variant were diagnosed before the age of 65, implicating that the
285 gene is involved in both early- and late-onset AD.

286 Holstege, et al. [23], reported that strongly damaging, but rare variants (ExAC[12] MAF < $1x10^{-5}$) in
287 *SORL1* as defined by a Combined Annotation Dependent Depletion (CADD) score of greater
288 than 30, increased the risk of Alzheimer's disease by 12-fold. The authors proposed that the
289 presence of these variants should be considered in addition to risk variants in *APOE,* and
290 causal variants in *PSEN1, PSEN2* or *APP* for assessing risk in a clinical setting. Accordingly,
291 only one of the *SORL1* variants identified in our study was found in ExAC[12], and was very rare
292 (11:121440980; ExAC AF = $4.95x10^{-5}$). Furthermore, half of the 10 variants with a CADD score
293 available were over 30, and all were over 25. The depletion of loss-of-function variants in the
294 ExAC database lends further evidence to the significance of the higher frequency of loss-of-
295 function variants in our AD sample.
296

297 *SORL1*, also known as *SORLA* and *LR11*, encodes a trafficking protein (sortilin-related
298 receptor, L(DLR class) A repeats containing protein) that binds the amyloid precursor protein
299 (APP) redirecting it to a non-amyloidogenic pathway within the retromer complex. The major site
300 for expression of SORL1 protein is in the brain especially within neurons and astrocytes. Aβ
301 peptides are also directed to the lysosome by SORL1. Processing of APP requires endocytosis
302 of molecules from the cell surface to endosomes whereby proteolytic breakdown to Aβ occurs.
303 SORL1 acts as a sorting receptor for APP that recycles molecules from endosomes back to the
304 trans-Golgi network to decrease Aβ production. We found that in the absence of the SORL1
305 gene, APP was released into the late endosome where it underwent β-secretase and γ-
306 secretase cleavage generating Aβ [24]. Thus, the mechanisms by which mutations in SORL1 lead
307 to neurodegeration in Alzheimer's disease relates to the disruption of its ability to bind APP.

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

308 Qualifying variants in other genes were also more prevalent among patients with AD compared
309 with healthy, non-demented controls. Variants in *GRID2IP*, *WDR76* and *GRN* were four to five
310 times more frequent in cases than in controls, though these genes have not yet achieved
311 genome-wide significance and thus further studies including larger patient samples will help
312 determine which contribute to AD risk.

313 Glutamate receptor delta-2 interacting protein (*GRID2IP*) is selectively expressed in the
314 cerebellar Purkinje cell-fiber synapses. The exact role for this gene is not fully understood, but it
315 appears to be a postsynaptic scaffold protein that links to GRID2 with signaling molecules and
316 the actin cytoskeleton [27]. There is no known role for *GRID2IP* in AD despite the fact that
317 mutations were found in two individuals with postmortem confirmed Alzheimer's disease. The
318 gene has not been well represented in existing exome sequencing libraries and the resulting
319 reduced coverage of this gene makes the findings more difficult to interpret. However, the
320 variants driving the signal in our analyses are all well covered in our entire cohort, with more
321 than 96% of samples achieving at least 10X coverage.

322 *WDR76* interacts with chromatin components and the cytosolic chaperonin containing TCP-1
323 (CCT), allowing for the maintenance of cellular homeostasis by assisting in the identification of
324 folded proteins. *WDR76* has low expression in brain and relatively high expression in lymph
325 nodes. Only one of the three individuals with postmortem data met "high likelihood criteria" for
326 AD.

327 *GRN* mutations in patients with clinically diagnosed AD have been previously reported in large
328 families in the National Institute on Aging family-based study (NIA-AD) [28] and among large,
329 multiply affected families of Caribbean Hispanic ancestry [29]. These loss-of-function mutations
330 result in haploinsufficiency, premature stop codons or nonsense variants impairing the secretion
331 or the structure of Progranulin, involved intracellular trafficking and lysosomal biogenesis and

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

332 function. Its role in AD in unclear and possibly coincidental [30]. The phenotype of FTLD includes
333 unique manifestations allowing it to be distinguished from AD. A family presumed to have
334 Alzheimer's disease phenotypically with a *GRN* mutation (c.154delA) had FTLD with ubiquitin-
335 positive, tau-negative and lentiform neuronal intranuclear inclusions (-U NII) with neuronal loss
336 and gliosis affecting the frontal and temporal lobes, and TDP43 inclusions [31]. Only one of the six
337 family members (Patient II:1) had mixed pathology meeting NIA-Reagan criteria of high
338 likelihood [16] and coexisting FTLD-U N11 with TDP43 inclusions. *GRN* mutations were also
339 observed in a sporadic patient with postmortem evidence of Alzheimer's disease: NIA-Reagan
340 criteria of high likelihood[16] and coexisting FTLD-U N11 with TDP43 inclusions [32]. Among the
341 patients with *GRN* mutations in this study, one patient met criteria for definite Alzheimer's
342 disease without co-existing FTLD, while another met pathological criteria for FTLD.

343 The results here indicate that extremely rare, loss-of-function variants in *SORL1* have an
344 strongly effect the risk of sporadic AD. While qualifying variants were present in only 0.27% of
345 patients, only a single variant was found among 13,232 controls, and the single control carrier
346 upon a post hoc cognitive evaluation was identified to have a diagnosis of mild cognitive
347 impairment. These results confirm and greatly extend those from sequencing studies in familial
348 and sporadic early onset Alzheimer's disease [19-21], familial AD families [24, 26, 33] and investigations
349 within clinical settings. The resulting impact of the loss-of-function variants in *SORL1* on
350 recycling of the amyloid precursor protein and the amyloid β protein make this pathway an
351 attractive target for the development of therapies. Beyond implicating *SORL1* and highly
352 suggestive candidate genes for AD, this study shows for the first time that the collapsing
353 analysis methodology of ultra-rare variants described here that has proven successful for a
354 number of rare diseases also can securely implicate genes in a condition as common as AD.

355

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

356 **Author Contributions**
357 Study Design:
358 NSR, CW, SK, SP, GT, BNV, DBG, RM
359 Data Collection:
360 AMB, HA, JJM, NS, RL, CW, SK, SP, GT, BNV, DBG, RM
361 Data Analysis:
362 NSR, CW, SK, SP, GT, BNV, DBG, RM
363 Writing and Editing:
364 NSR, AMB, HA, JJM, NS, RL, CW, SK, SP, GT, BNV, DBG, RM

365 **Acknowledgements**

366 WHICAP and EFIGA
367 Data collection for this project was supported by the Washington Heights and Inwood
368 Community Aging Project (WHICAP) and Genetic Studies of Alzheimer's disease in Caribbean
369 Hispanics (Estudio familiar de la genética de la enfermedad de Alzheimer, also known as
370 EFIGA) funded by the National Institute on Aging (NIA), by the National Institutes of Health
371 (NIH) (1RF1AG054023, 5R37AG015473, RF1AG015473, R56AG051876), and the National
372 Center for Advancing Translational Sciences, NIH through Grant Number TL1TR001875. We
373 acknowledge the WHICAP and EFIGA study participants and the research and support staff for
374 their contributions to this study.

375 ADSP
376 The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease
377 (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded
378 Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the
379 Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

380 Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA
381 (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute
382 of Health (NIH) institutes and other foreign governmental and non-governmental organizations.
383 The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs.
384 Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri;
385 U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr.
386 Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr.
387 Goate and U01AG052410 to Dr. Pericak-Vance. Data generation and harmonization in the
388 Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang).

389 The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer's Disease Centers
390 (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP),
391 Mayo Clinic (MAYO), Mayo Parkinson's Disease controls, University of Miami, the Multi-
392 Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE), the National Cell
393 Repository for Alzheimer's Disease (NCRAD), the National Institute on Aging Late Onset
394 Alzheimer's Disease Family Study (NIA-AD; U24 AG056270), the Religious Orders Study
395 (ROS), the Texas Alzheimer's Research and Care Consortium (TARC), Vanderbilt
396 University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood
397 Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP),
398 the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer
399 (EFIGA), the University of Toronto (UT), and Genetic Differences (GD).
400

401 The CHARGE cohorts, with funding provided by 5RC2HL102419 and HL105756, include the
402 following: Atherosclerosis Risk in Communities (ARIC) Study which is carried out as a
403 collaborative study supported by NHLBI contracts (HHSN268201100005C,
404 HHSN268201100006C, HHSN268201100007C, HHSN268201100008C,

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

405 HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and
406 HHSN268201100012C), Austrian Stroke Prevention Study (ASPS), Cardiovascular Health
407 Study (CHS), Erasmus Rucphen Family Study (ERF), Framingham Heart Study (FHS), and
408 Rotterdam Study (RS). CHS research was supported by contracts HHSN268201200036C,
409 HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081,
410 N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from
411 the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the
412 National Institute of Neurological Disorders and Stroke (NINDS). Additional support was
413 provided by R01AG023629, R01AG15928, and R01AG20098 from the National Institute on
414 Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-
415 NHLBI.org. The content is solely the responsibility of the authors and does not necessarily
416 represent the official views of the National Institutes of Health.
417 The three LSACs are: the Human Genome Sequencing Center at the Baylor College of
418 Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), and the
419 Washington University Genome Institute (U54HG003079).
420

421 Biological samples and associated phenotypic data used in primary data analyses were stored
422 at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease
423 (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data
424 used in primary and secondary data analyses were provided by Study Investigators, the NIA
425 funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center
426 (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease
427 Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA,
428 and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was
429 supported in part by the Intramural Research Program of the National Institutes of health,
430 National Library of Medicine. Contributors to the Genetic Analysis Data included Study

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

431 Investigators on projects that were individually funded by NIA, and other NIH institutes, and by
432 private U.S. organizations, or foreign governmental or nongovernmental organizations.
433

434 We would like to acknowledge the following individuals or groups for the contributions of control
435 samples: T. Young; K. Whisenhunt; S. Palmer; S. Berkovic, I. Scheffer, B. Grinton; E. Cirulli; M.
436 Winn; R.Gbadegesin; A. Poduri; S. Schuman; E. Nading; E. Pras; D. Lancet; Z. Farfel; S.
437 Kerns; H. Oster; D. Valle; J. Hoover-Fong; N. Sobriera; M. Hauser; G. Nestadt; J. Samuels; Y.
438 Wang; G. Cavalleri, N. Delanty; C. Depondt; S. Sisodiya; R. Buckley; C. Moylan; A. M. Diehl;
439 M. Abdelmalek; S. Delaney; V. Shashi; M. Walker; M. Sum; the ALS Sequencing Consortium;
440 the Washington University Neuromuscular Genetics Project; DUHS (Duke University Health
441 System) Nonalcoholic Fatty Liver Disease Research Database and Specimen Repository;
442 Epilepsy Genetics Initiative, A Signature Program of CURE; the Epi4K Consortium and Epilepsy
443 Phenome/Genome Project; the Undiagnosed Diseases Network; and the Utah Foundation for
444 Biomedical Research.

445 The collection of control samples and data was funded in part by: Biogen; Gilead Sciences, Inc.;
446 UCB; National Institutes of Health (RO1HD048805); National Institute of Neurological Disorders
447 and Stroke (U01NS077303, U01NS053998, U54NS078059); National Institute of Child Health
448 and Human Development (P01HD080642); National Institute of Mental Health (R01MH097971,
449 K01MH098126); National Human Genome Research Institute (U01HG007672); an American
450 Academy of Child and Adolescent Psychiatry (AACAP) Pilot Research Award; Endocrine
451 Fellows Foundation Grant; the NIH Clinical and Translational Science Award Program
452 (UL1TR000040); the Ellison Medical Foundation New Scholar award AG-NS-0441-08; Duke
453 Chancellor's Discovery Program Research Fund 2014; The J. Willard and Alice S. Marriott
454 Foundation; The Muscular Dystrophy Association; The Nicholas Nunno Foundation; The JDM
455 Fund for Mitochondrial Research; The Arturo Estopinan TK2 Research Fund; the Stanley

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

456 Institute for Cognitive Genomics at Cold Spring Harbor Laboratory; New York-Presbyterian
457 Hospital; the Columbia University College of Physicians and Surgeons; and the Columbia
458 University Medical Center.

459 The content is solely the responsibility of the authors and does not necessarily represent the
460 official views of the National Institutes of Health.

461

462 Biogen Inc. provided support for whole exome sequencing for the WHICAP cohort through a
463 grant to David Goldstein, PhD and salary support for Neha S. Raghavan PhD for analyses.
464 Individuals at Biogen were not involved in the collection of data, analysis or interpretation of the
465 genetic data, nor in the production of this manuscript.

466 **Declaration of interests**

467 SP is a paid employee of and holds stock in AstraZeneca. All other authors have no interests to
468 declare.

469

470

471

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

472 **References**

473 1. Consortium UK, Walter K, Min JL, et al. The UK10K project identifies rare variants in
474 health and disease. Nature. 2015 Oct 1;526(7571):82-90.
475 2. Keinan A, Clark AG. Recent explosive human population growth has resulted in an
476 excess of rare genetic variants. Science. 2012 May 11;336(6082):740-3.
477 3. Park JH, Gail MH, Weinberg CR, et al. Distribution of allele frequencies and effect sizes
478 and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S
479 A. 2011 Nov 1;108(44):18026-31.
480 4. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and
481 opportunities. Genome Med. 2015;7(1):16.
482 5. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral
483 sclerosis identifies risk genes and pathways. Science. 2015 Mar 27;347(6229):1436-41.
484 6. Epi Kc, Epilepsy Phenome/Genome P. Ultra-rare genetic variation in common epilepsies:
485 a case-control sequencing study. Lancet Neurol. 2017 Feb;16(2):135-43.
486 7. Beecham GW, Bis JC, Martin ER, et al. The Alzheimer's Disease Sequencing Project:
487 Study design and sample selection. Neurol Genet. 2017 Oct;3(5):e194.
488 8. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical
489 diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the
490 auspices of Department of Health and Human Services Task Force on Alzheimer's Disease.
491 Neurology. 1984 Jul;34(7):939-44.
492 9. Kawas C, Segal J, Stewart WF, Corrada M, Thal LJ. A validation study of the Dementia
493 Questionnaire. Arch Neurol. 1994 Sep;51(9):901-6.
494 10. Brandt JS, M.; Folstein, M. The telephone interview for cognitive status. Neuropsychiatry
495 Neuropsychol Behav Neurol July 1988;1(2):111-7.
496 11. Farrell CM, O'Leary NA, Harte RA, et al. Current status and new features of the
497 Consensus Coding Sequence database. Nucleic acids research. 2014 Jan;42(Database
498 issue):D865-72.
499 12. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in
500 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91.
501 13. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals
502 identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013 Dec;45(12):1452-
503 8.
504 14. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide
505 association scans. Bioinformatics. 2010 Sep 1;26(17):2190-1.
506 15. Albert MS, DeKosky ST, Dickson D, et al. The diagnosis of mild cognitive impairment
507 due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's
508 Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement.
509 2011 May;7(3):270-9.
510 16. Newell KL, Hyman BT, Growdon JH, Hedley-Whyte ET. Application of the National
511 Institute on Aging (NIA)-Reagan Institute criteria for the neuropathological diagnosis of
512 Alzheimer disease. J Neuropathol Exp Neurol. 1999 Nov;58(11):1147-55.
513 17. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging
514 missense mutations. Nat Methods. 2010 Apr;7(4):248-9.

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

515 18. Cairns NJ, Bigio EH, Mackenzie IR, et al. Neuropathologic diagnostic and nosologic
516 criteria for frontotemporal lobar degeneration: consensus of the Consortium for Frontotemporal
517 Lobar Degeneration. Acta Neuropathol. 2007 Jul;114(1):5-22.
518 19. Pottier C, Hannequin D, Coutant S, et al. High frequency of potentially pathogenic
519 SORL1 mutations in autosomal dominant early-onset Alzheimer disease. Mol Psychiatr. 2012
520 Sep;17(9):875-9.
521 20. Nicolas G, Charbonnier C, Wallon D, et al. SORL1 rare variants: a major risk factor for
522 familial early-onset Alzheimer's disease. Mol Psychiatry. 2016 Jun;21(6):831-6.
523 21. Verheijen J, Van den Bossche T, van der Zee J, et al. A comprehensive study of the
524 genetic impact of rare variants in SORL1 in European early-onset Alzheimer's disease. Acta
525 Neuropathol. 2016 Aug;132(2):213-24.
526 22. Cuccaro ML, Carney RM, Zhang Y, et al. SORL1 mutations in early- and late-onset
527 Alzheimer disease. Neurol Genet. 2016 Dec;2(6):e116.
528 23. Holstege H, van der Lee SJ, Hulsman M, et al. Characterization of pathogenic SORL1
529 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy.
530 European journal of human genetics : EJHG. 2017 Aug;25(8):973-81.
531 24. Rogaeva E, Meng Y, Lee JH, et al. The neuronal sortilin-related receptor SORL1 is
532 genetically associated with Alzheimer disease. Nat Genet. 2007 Feb;39(2):168-77.
533 25. Bellenguez C, Charbonnier C, Grenier-Boley B, et al. Contribution to Alzheimer's
534 disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls.
535 Neurobiol Aging. 2017 Nov;59:220 e1- e9.
536 26. Vardarajan BN, Zhang Y, Lee JH, et al. Coding mutations in SORL1 and Alzheimer
537 disease. Ann Neurol. 2015 Feb;77(2):215-27.
538 27. Sonoda T, Mochizuki C, Yamashita T, et al. Binding of glutamate receptor delta2 to its
539 scaffold protein, Delphilin, is regulated by PKA. Biochem Biophys Res Commun. 2006 Nov
540 24;350(3):748-52.
541 28. Cruchaga C, Haller G, Chakraverty S, et al. Rare variants in APP, PSEN1 and PSEN2
542 increase risk for AD in late-onset Alzheimer's disease families. PLoS One. 2012;7(2):e31039.
543 29. Lee JH, Kahn A, Cheng R, et al. Disease-related mutations among Caribbean Hispanics
544 with familial dementia. Mol Genet Genomic Med. 2014 Sep;2(5):430-7.
545 30. Kao AW, McKay A, Singh PP, Brunet A, Huang EJ. Progranulin, lysosomal regulation
546 and neurodegenerative disease. Nat Rev Neurosci. 2017 Jun;18(6):325-33.
547 31. Kelley BJ, Haidar W, Boeve BF, et al. Alzheimer disease-like phenotype associated with
548 the c.154delA mutation in progranulin. Arch Neurol. 2010 Feb;67(2):171-7.
549 32. Perry DC, Lehmann M, Yokoyama JS, et al. Progranulin mutations as risk factors for
550 Alzheimer disease. JAMA Neurol. 2013 Jun;70(6):774-8.
551 33. Lee JH, Cheng R, Schupf N, et al. The association between genetic variants in SORL1
552 and Alzheimer disease in an urban, multiethnic, community-based cohort. Arch Neurol. 2007
553 Apr;64(4):501-6.

554

555

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

556 **Figure Legend**

557

558 Figure 1. QQ Plot: Observed vs. expected p-values. Lambda = 1.04173

559

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

560

Table 1. Characteristics of Study Cohort (n=20,197)

	AD Cases		Controls		
	WHICAP	**ADSP**	**WHICAP**	**ADSP**	**External**
N	1371	5594	2331	4506	6395
Combined	**6965**		**13,232**		
Age (mean ± SD)	81.4±6.2	75.4±8.4	78.1±6.8	86.07±4.53	N/A
Combined	**76.7±8.5**		**83.4±6.7**		N/A
Sex (%F)	68.5	57.2	67.6	41.1	47.3
Combined	**59.4**		**45.2**		
APOE E4 (% Carrier)	27.43	42.40	20.94	15.14	N/A
Combined	**39.50**		**17.10**		N/A

Mean age and APOE E4 carrier % do not include the External controls;
Age for cases indicates age at diagnosis, and for controls the age at last assessment or age when last known to be free of dementia

561

562

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

563

Table 2. Variant counts for the top four AD genes

Gene Name	Total Variant	Total SNV	Total Indel	No. of Cases w/ QV	Case Frequency	No. of Cntrls w/ QV	Control Frequency	Enriched Direction	Fet P
SORL1	17	10	7	19	0.0027	1	7.56E-05	case	2.17E-08
GRID2IP	12	5	8	11	0.0016	2	1.51E-04	case	2.98E-04
WDR76	10	3	7	10	0.0014	2	1.51E-04	case	7.39E-04
GRN	12	6	6	11	0.0016	3	2.27E-04	case	9.56E-04

564 *QV= Qualifying variant; FET = Fisher's Exact Test*

565

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Table 3. SORL1 variants

Genomic Position	Variant Type	Variant Class	CADD score	Protein modification	ExAC Global Frequency	Case/ Control	Sex	Ethnicity	Braak Stage	Age at Onset or Last Visit
11-121367577	snv	SAV	26.6	NA	0	case	F	AA	NA	77
11-121367654	snv	SG	37	p.Arg279*	0	case	F	NHW	6	72
11-12142134322[23]	snv	SG	39	p.Arg744*	0	case	M	NHW	NA	65
11-12142134322[23]	snv	SG	39	p.Arg744*	0	case	F	NHW	NA	67
11-121426001	indel	FV	NA	p.Asp850fs	0	case	F	NHW	NA	60
11-121428047	snv	SG	41	p.Arg866*	0	case	M	NHW	6	65
11-121430263	indel	FV	NA	p.Ile983fs	0	ctrl	M	AA	NA	64
11-121440980	snv	SDV	27.6	NA	4.95E-05	case	F	CH	NA	80
11-121456930	snv	SAV	26.8	NA	0	case	M	NHW	NA	69
11-121456930	snv	SAV	26.8	NA	0	case	M	NHW	6	62
11-121461788	indel	FV	NA	p.Cys1431fs	0	case	F	NHW	NA	61
11-12146648224[25]	snv	SDV	28	NA	0	case	F	NHW	3	90+
11-12146648224[25]	snv	SDV	28	NA	0	case	F	NHW	NA	90+
11-121474911	indel	FV	NA	p.Thr1511fs	0	case	M	NHW	NA	60
11-121474984	snv	SG	35	p.Cys1534*	0	case	F	NHW	NA	74
11-12147756824[25]	snv	SG	46	p.Arg1655*	0	case	M	NHW	NA	69
11-121477667	snv	SDV	26.9	NA	0	case	F	AA	NA	68
11-121485637	indel	FV	NA	p.Asp1828fs	0	case	M	NHW	NA	75
11-121491801	indel	FV	NA	p.Lys1975fs	0	case	M	NHW	6	61
11-121500253	indel	FV	NA	p.Met2211fs	0	case	M	NHW	6	62

566

567 Those in bold have previously been identified as indicated by the reference

568 SNV = Single Nucelotide Variant; Indel = Insertion or Deletion

569 CADD = Combined Annotation Dependent Depletion

570 FV = Frameshift Variant; SAV = Splice Acceptor Variant; SDV = Splice Donor Variant; SG = Stop Gained

571 AA = African American; CH = Carribean Hispanic; NHW = Non-hispanic White

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

572
573
574 **Table 4. Counts of ultra-rare variant in previously identified or implicated AD genes**
575

Gene Name	Cases w/ QV	Cases w/o QV	Controls w/ QV	Controls w/o QV	FET p-value
ABCA7	28	6937	34	13198	0.08
APOE	0	6965	2	13230	0.55
APP	2	6963	2	13230	0.61
BIN1	1	6964	2	13230	1.00
CASS4	1	6964	1	13231	1.00
CD2AP	0	6965	6	13226	0.10
CELF1	1	6964	0	13232	0.34
CLU	1	6964	1	13231	1.00
CR1	6	6959	17	13215	0.65
EPHA1	6	6959	23	13209	0.17
FERMT2	0	6965	1	13231	1.00
HLA-DRB5	9	6956	12	13220	0.46
INPP5D	1	6964	1	13231	1.00
MEF2C	1	6964	3	13229	1.00
MS4A6A	2	6963	7	13225	0.72
NME8	11	6954	11	13221	0.18
PICALM	1	6964	3	13229	1.00
PSEN1	2	6963	0	13232	0.12
PSEN2	2	6963	0	13232	0.12
PTK2B	6	6959	10	13222	0.80
SLC24A4	1	6964	3	13229	1.00
SORL1	19	6946	1	13231	2.17E-08
TREM2	4	6961	4	13228	0.46
ZCWPW1	9	6956	5	13227	0.02
Total	**114**	**6857**	**149**	**13087**	

	Cases	Controls
Total % w/ variant	1.6	1.1
Total FET p-val	0.002	

576
577 Qualifying loss-of-function variants per gene and combined across the 24 genes; QV = Qualifying
578 variant, FET = Fisher's exact test
579

bioRxiv preprint first posted online Apr. 20, 2018; doi: http://dx.doi.org/10.1101/305631. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.