Exome sequencing identifies rare damaging variants in *ATP8B4* and *ABCA1* as novel risk factors for Alzheimer's Disease

Authors:

Henne Holstege[1,2,3,*,#]; Marc Hulsman[1,2,3,*,#]; Camille Charbonnier[4,*]; Benjamin Grenier-Boley[5]; Olivier Quenez[4]; Detelina Grozeva[6]; Jeroen G.J. van Rooij[7,8]; Rebecca Sims[6]; Shahzad Ahmad[9,10]; Najaf Amin[9,11]; Penny J. Norsworthy[12]; Oriol Dols-Icardo[13,14]; Holger Hummerich[12]; Amit Kawalia[15]; Alzheimer's Disease Neuroimaging Initiative (ADNI) database[16]; Philippe Amouyel[5]; Gary W. Beecham[17]; Claudine Berr[18]; Joshua C. Bis[19]; Anne Boland[20]; Paola Bossù[21]; Femke Bouwman[2]; Jose Bras[22,23]; Dominique Campion[4]; J. Nicholas Cochran[24]; Antonio Daniele[25]; Jean-François Dartigues[26]; Stéphanie Debette[26,27]; Jean-François Deleuze[20]; Nicola Denning[28]; Anita L DeStefano[29,30,31]; Lindsay A. Farrer[29,31,32,33]; Maria Victoria Fernandez[34,35,36]; Nick C. Fox[37]; Daniela Galimberti[38,39]; Emmanuelle Genin[40]; Hans Gille[41]; Yann Le Guen[42]; Rita Guerreiro[22,23]; Jonathan L. Haines[43]; Clive Holmes[44]; M. Arfan Ikram[9]; M. Kamran Ikram[9]; Iris E. Jansen[2,45]; Robert Kraaij[8]; Marc Lathrop[46]; Afina W. Lemstra[2]; Alberto Lleó[13,14]; Lauren Luckcuck[6]; Marcel M. A. M. Mannens[47]; Rachel Marshall[6]; Eden R Martin[17,48]; Carlo Masullo[49]; Richard Mayeux[50,51]; Patrizia Mecocci[52]; Alun Meggy[28]; Merel O. Mol[7]; Kevin Morgan[53]; Richard M. Myers[24]; Benedetta Nacmias[54,55]; Adam C Naj[56,57]; Valerio Napolioni[42,58]; Florence Pasquier[59]; Pau Pastor[60,61]; Margaret A. Pericak-Vance[17,48]; Rachel Raybould[28]; Richard Redon[62]; Marcel J.T. Reinders[3]; Anne-Claire Richard[4]; Steffi G Riedel-Heller[63]; Fernando Rivadeneira[8]; Stéphane Rousseau[4]; Natalie S. Ryan[37]; Salha Saad[6]; Pascual Sanchez-Juan[14,64]; Gerard D. Schellenberg[57]; Philip Scheltens[2]; Jonathan M. Schott[37]; Davide Seripa[65]; Sudha Seshadri[30,31,66]; Daoud Sie[41]; Erik Sistermans[41]; Sandro Sorbi[54,55]; Resie van Spaendonk[41]; Gianfranco Spalletta[67]; Niccólo Tesi[1,2,3]; Betty Tijms[2]; André G Uitterlinden[8]; Sven J. van der Lee[1,2,3]; Pieter Jelle de Visser[2]; Michael Wagner[68,69]; David Wallon[70]; Li-San Wang[57]; Aline Zarea[70]; Jordi Clarimon[13,14]; John C. van Swieten[7]; Michael D. Greicius[42]; Jennifer S. Yokoyama[72]; Carlos Cruchaga[34,35,36]; John Hardy[71]; Alfredo Ramirez[15,66,68,69,73]; Simon Mead[12]; Wiesje M. van der Flier[2]; Cornelia M van Duijn[9,11]; Julie Williams[6]; Gaël Nicolas[4,#,*]; Céline Bellenguez[5,*]; Jean-Charles Lambert[5,#,*]

*Authors contributed equally to this work

[#] To whom correspondence should be addressed

— Henne Holstege: h.holstege@amsterdamumc.nl
— Marc Hulsman: m.hulsman@amsterdamumc.nl
— Gael Nicolas: gaelnicolas@hotmail.com
— Jean-Charles Lambert: jean-charles.lambert@pasteur-lille.fr

Affiliations

1. Genomics of Neurodegenerative Diseases and Aging, Department of Human Genetics, Amsterdam University Medical Center (location VUmc), Neuroscience Campus Amsterdam, Amsterdam, The Netherlands
2. Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
3. Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
4. Normandie Univ, UNIROUEN, Inserm U1245 and CHU Rouen, Department of Genetics and CNRMAJ, F-76000 Rouen, France
5. Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1167-RID-AGE facteurs de risque et déterminants moléculaires des maladies liés au vieillissement, Lille, France
6. MRC Centre for Neuropsychiatric Genetics and Genomics, , Division of Psychological Medicine and Clinical Neuroscience, School of Medicine, Cardiff University, Cardiff, UK
7. Department of Neurology, Erasmus Medical Centre, Rotterdam, The Netherlands
8. Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, The Netherlands
9. Department of Epidemiology, Erasmus Medical Centre, Rotterdam, The Netherlands
10. LACDR, Leiden, The Netherlands
11. Nuffield Department of Population Health Oxford University
12. MRC Prion Unit at UCL, UCL Institute of Prion Diseases, London, UK
13. Department of Neurology, II B Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain
14. CIBERNED, Network Center for Biomedical Research in Neurodegenerative Diseases, National Institute of Health Carlos III, Madrid, Spain
15. Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne Germany
16. ADNI consortium; see supplemental authors
17. The John P. Hussman Institute for Human Genomics, University of Miami, Miami, Florida, USA
18. Univ Montpellier, Inserm, INM (Institute for Neurosciences of Montpellier), Montpellier, France
19. Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA.
20. Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine Evry, France
21. Experimental Neuro-psychobiology Laboratory, Department of Clinical and Behavioral Neurology, IRCCS Santa Lucia Foundation, Rome, Italy
22. Department of Neurodegenerative Science, Van Andel Institute, Grand Rapids, MI USA
23. Division of Psychiatry and Behavioral Medicine, Michigan State University College of Human Medicine, Grand Rapids, MI, USA.
24. HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA
25. Department of Neuroscience, Catholic University of Sacred Heart, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
26. University Bordeaux, Inserm, Bordeaux Population Health Research Center, France
27. Department of Neurology, Bordeaux University Hospital, Bordeaux, France
28. UKDRI@ Cardiff, School of Medicine, Cardiff University, Cardiff, UK
29. Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.
30. Framingham Heart Study, Framingham, MA, USA.
31. Department of Neurology, Boston University School of Medicine, Boston, MA, USA.
32. Department of Epidemiology, Boston University, Boston, MA, USA

33. Department of Medicine (Biomedical Genetics), Boston University, Boston, MA, USA
34. Neurogenomics and Informatics Center, Washington University School of Medicine, St Louis, MO USA
35. Psychiatry Department, Washington University School of Medicine, St Louis, MO
36. Hope Center for Neurological Disorders, Washington University School of Medicine, St Louis, MO
37. Dementia Research Centre, UCL Queen Square Institute of Neurology, London, UK
38. Fondazione IRCCS Ca' Granda, Ospedale Policlinico, Milan, Italy
39. University of Milan, Milan, Italy
40. Univ Brest, Inserm, EFS, CHU Brest, UMR 1078, GGB, F-29200, Brest, France
41. Genome Diagnostics, Department of Human Genetics, VU University, AmsterdamUMC (location VUmc), Van Der Boechorststraat 7, 1081 BT Amsterdam
42. Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA.
43. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA
44. Clinical and Experimental Science, Faculty of Medicine, University of Southampton, Southampton, UK.
45. Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije University, Amsterdam, The Netherlands.
46. McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada
47. Department of Human Genetics, Amsterdam UMC, University of Amsterdam, Amsterdam Reproduction and Development Research Institute Amsterdam
48. Dr. John T. Macdonald Foundation Department of Human Genetics, University of Miami, Miami, Florida, USA
49. Institute of Neurology, Catholic University of the Sacred Heart , Rome, Italy
50. Taub Institute on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University, New York, New York, USA
51. Gertrude H. Sergievsky Center, Columbia University, New York, New York,
52. Institute of Gerontology and Geriatrics, Department of Medicine and Surgery, University of Perugia, Perugia, Italy
53. Human Genetics, School of Life Sciences, University of Nottingham, UK
54. Department of Neuroscience, Psychology, Drug Research and Child Health University of Florence, Florence Italy
55. IRCCS Fondazione Don Carlo Gnocchi, Florence, Italy
56. Penn Neurodegeneration Genomics Center, Department of Biostatistics, Epidemiology, and Informatics; University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania,
57. Penn Neurodegeneration Genomics Center, Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania,
58. Genomic And Molecular Epidemiology (GAME) Lab, School of Biosciences and Veterinary Medicine, University of Camerino (UNICAM) Camerino, 62032, Italy
59. Univ. Lille, Inserm, CHU Lille, UMR1172, Resources and Research Memory Center (MRRC) of Distalz, Licend, Lille France
60. Fundació Docència i Recerca MútuaTerrassa and Movement Disorders Unit, Department of Neurology, University Hospital MútuaTerrassa, Barcelona, Spain
61. Memory Disorders Unit, Department of Neurology, Hospital Universitari Mutua de Terrassa, Terrassa, Barcelona, Spain.
62. Université de Nantes, CHU Nantes, CNRS, INSERM, l'institut du thorax, Nantes, France
63. Institute of Social Medicine, Occupational Health and Public Health, University of Leipzig, Leipzig, Germany.
64. Neurology Service, Marqués de Valdecilla University Hospital (University of Cantabria and IDIVAL), Santander, Spain.

65. Laboratory for Advanced Hematological Diagnostics, Department of Hematology and Stem Cell Transplant, Lecce, Italy
66. Department of Psychiatry and Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, San Antonio, TX, USA
67. Laboratory of Neuropsychiatry, Department of Clinical and Behavioral Neurology, IRCCS Santa Lucia Foundation, Rome, Italy
68. Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Medical Faculty, Bonn, Germany.
69. German Center for Neurodegenerative Diseases (DZNE, Bonn), Bonn, Germany.
70. Normandie Univ, UNIROUEN, Inserm U1245 and CHU Rouen, Department of Neurology and CNRMAJ, F-76000 Rouen, France
71. Reta Lila Weston Research Laboratories, Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK.
72. Memory and Aging Center, Department of Neurology, University of California, San Francisco, CA USA
73. Cluster of Excellence Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Cologne, Germany

The genetic component of Alzheimer's disease (AD) has been mainly assessed using Genome Wide Association Studies (GWAS), which do not capture the risk contributed by rare variants. Here, we compared the gene-based burden of rare damaging variants in exome sequencing data from 32,558 individuals —16,036 AD cases and 16,522 controls— in a two-stage analysis. Next to known genes *TREM2*, *SORL1* and *ABCA7*, we observed a significant association of rare, predicted damaging variants in *ATP8B4* and *ABCA1* with AD risk, and a suggestive signal in *ADAM10*. Next to these genes, the rare variant burden in *RIN3, CLU, ZCWPW1* and *ACE* highlighted these genes as potential driver genes in AD-GWAS loci. Rare damaging variants in these genes, and in particular loss-of-function variants, have a large effect on AD-risk, and they are enriched in early onset AD cases. The newly identified AD-associated genes provide additional evidence for a major role for APP-processing, Aβ-aggregation, lipid metabolism and microglial function in AD.

AD is the leading cause of dementia and its impact will grow with increasing life expectancy[1]. Beyond autosomal dominant early onset AD (<1% of all AD cases, onset ≤65 years), the common complex form of AD has an estimated heritability of ~70%[2]. Using genome wide association studies (GWAS), 75 mostly common genetic risk factors/loci have been associated with AD-risk in populations with European ancestry, but individually these common variants have low effect-sizes[3]. Using DNA-sequencing strategies, rare (allele frequency <1%) damaging missense or loss-of-function (LOF) variants in the *TREM2*, *SORL1* and *ABCA7* genes were identified to also contribute to the heritability of AD, with substantially higher effect-sizes than individual GWAS-hits[4-8].

In order to detect additional genes for which rare variants associate with AD-risk, it is necessary to compare genetic sequencing data from thousands of AD cases and controls. In a large collaborative effort, we harmonized sequencing data from studies from Europe and the United States (**Table S1**) and applied a multi-stage gene burden analysis (**Figure 1A**). We observed site-specific technical biases, since data were generated at multiple centers, using heterogeneous methods. To account for these

batch effects, we designed and applied comprehensive quality control (QC) procedures (**Online methods, Tables S2-S3**).

After sample QC, we first compared gene-based rare-variant burdens between 12,652 AD cases and 8,693 controls (Stage-1 analysis, **Table 1A**). We detected 7,543,193 variants after sample- and variant-QC (**Table 1B**) and annotated LOF variants with LOFTEE and missense variants with the REVEL score, and selected variants with MAF<1%. We defined four deleteriousness thresholds by incrementally including variants with lower levels of predicted deleteriousness: respectively LOF (n=57,543), LOF+REVEL≥75 (n=111,755), LOF+REVEL≥50 (n=211,665), and LOF+REVEL≥25 (n=409,733). Of the 19,822 autosomal protein coding genes, we analyzed the 13,222 genes that had a cumulative minor allele count (cMAC) ≥10 for the lowest deleterious threshold LOF+REVEL≥25 (**see Methods**); 9,168 genes for the LOF+REVEL≥50 threshold; 5,694 for the LOF+REVEL≥75 threshold and 3,120 genes for the LOF-only threshold (**Figure 1B**). For these different deleteriousness thresholds, this analysis has an estimated power of 41%, 22%, 11% and 4%, respectively to attain a signal with p<1e-6, assuming that the differential variant burden for a gene is associated with an odds ratio of 10.0 in EOAD and 3.33 in LOAD (**Table S4**). Therefore, this analysis only has the power to uncover genes for which the differential gene-burden is associated with a large effect size or large numbers of damaging variant carriers (**Figure 1B**). In total, 31,204 tests were performed across 13,222 genes in Stage-1 (single genes were tested with up to four thresholds). Statistical inflation of test results was negligible (λ=1.046, **Figure 1C**). Of all burden tests performed, 13 tests, covering 6 genes indicated a suggestive differential variant burden between AD cases and controls (FDR<0.1): *SORL1*, *TREM2*, *ABCA7*, *ATP8B4*, *ADAM10*, and *ABCA1* (**Table 2A**).

To confirm these signals, we applied an analysis model consistent with Stage-1 to an independent Stage-2 dataset, which after QC, comprised 3,384 cases and 7,829 controls (**Table 1A**). The effect was tested in the direction observed in Stage-1 (one-sided test). All genes selected in Stage-1 reached p<0.05 (**Table 2A, Stage-2**). Stage-2 effect-sizes of these genes correlated with those observed in Stage-

1 (Pearson's r on log-odds: 0.91). We then meta-analyzed Stage-1 + Stage-2 across the 13 tests using a fixed-effect inverse variance method and corrected for the 31,204 tests performed in Stage-1 (Holm-Bonferroni) (**Table 2A**). This confirmed the AD-association of rare damaging variants in the *SORL1*, *TREM2*, *ABCA7*, *ATP8B4* and *ABCA1* genes. The association signal of the *ADAM10* gene was not exome-wide significant, presumably because prioritized variants in this gene are extremely few and rare, such that the signal can be confirmed only in larger datasets.

Strikingly, most of these genes also map to GWAS loci (*SORL1*, *TREM2*, *ABCA7*, *ABCA1* and *ADAM10)*. This led us to perform a focused analysis on GWAS loci, aiming to identify potential driver genes. To maximize statistical power, we merged the full exomes from the Stage-1 and Stage-2 samples into one mega-sample (**Table 1**). We interrogated genes that were previously prioritized to drive the AD association in the 75 loci identified in the most recent GWAS[3] (**Table S5, Online Methods**). In 67 genes, we observed sufficient prioritized variants (cMAC ≥10) to test the burden signal in at least one deleteriousness category (a total of 187 tests). In addition to the genes mentioned above, our analysis indicated a suggestive signal of increased AD risk in *RIN3, CLU, ZCWPW1,* and *ACE* (FDR<0.05) (**Table 2B, Table S6**); these signals will have to be confirmed in a larger dataset. Nevertheless, the AD associations in these genes persisted when focusing on the burden of only the very rare variants (MAF<0.1%), suggesting that the rare variant burden is not in linkage with, and thus independent from the GWAS sentinel variant.

Together, the newly associated genes provide additional evidence for a central role for *APP*-processing, lipid metabolism, Aβ-aggregation and neuroinflammatory processes in AD pathophysiology. Like *ABCA7*, *ATP8B4* is a phospholipid transporter. Rare variants in this gene have been associated with the risk of developing systemic sclerosis, an autoimmune disease[9]. In the brain, *ATP8B4* is predominantly expressed in microglia. Interestingly, GWAS indicated a potential association of *ATP8B4* with AD[3], mainly through the rare missense variant that was most recurrent in our study (G395S). Of note, the odds ratio point-estimate for *ATP8B4* LOF variants was close to 1, allowing for the possibility that the

missense variants that drive the *ATP8B4* association do not depend on a loss-of-function effect. ABCA1 is also a phospholipid transporter; it lipidates APOE[10] and poor ABCA1-dependent lipidation of APOE-containing lipoprotein particles increases Aβ-deposition and fibrillogenesis[11]. In line with this, the rare N1800H loss-of-function variant in *ABCA1* was previously associated with low plasma levels of ApoE and evidence suggested an association with increased risk of AD and cerebrovascular disease[12]. The α-secretase ADAM10 plays a major role in non-amyloidogenic APP metabolism[13]. Evidence for the AD-association of rare variants in *ADAM10* has remained suggestive until now: two rare missense variants in *ADAM10* were reported before to incompletely segregate with LOAD in a few families[14] (these variants did not associate with AD in our study, **Table SG6)** and a nonsense variant in the *ADAM10* gene was found to segregate with AD but in a small pedigree[15]. **Error! Bookmark not defined.***RIN3* has been associated with endosomal dysfunction and APP trafficking/metabolism[16,17]. *CLU* (also known as *APOJ*) has been found to affect Aβ-aggregation and clearance[18] and *ACE* is suggested to have a role in Aβ-degradation[19]. Thus far, the role of the histone methylation reader *ZCWPW1* remains unclear.

To better comprehend how these genes associate with AD, we analyzed the characteristics of rare damaging variants that contributed to the burden using the mega-sample (**Figure 2**, **Table 3**) For damaging variants in most genes, we observed increased carrier frequencies in younger cases and larger effect sizes were associated with an earlier age at onset (p=0.0001) (**Table S7**). Yet the variants also contributed to an increased risk of late-onset AD (**Figures 2A-B, Table 3**). The largest effect-sizes were measured for LOF variants in *SORL1, ADAM10*, *CLU* and *ZCWPW1,* and carriers of such variants had the lowest median age at onset (**Table 3**), implying a key role for these genes in AD etiology. Moderate variant-effect-sizes were observed for LOF variants in *TREM2*, *ABCA1* and *RIN3*, while the smallest variant-effects were observed in *ABCA7*, *ATP8B4* and *ACE (***Figure 3, Table 3**).

Extremely rare variants contributed more to large effect sizes than less rare variants (p=0.03, **Table S8**). Indeed, for *SORL1*, the variants with the lowest variant frequencies had the largest effect-sizes (**Figures 2C, Table S9**), and damaging variants in *ADAM10*, *CLU* and *ZCWPW1* were all extremely rare

(**Figures 2D**). Conversely, we observed that rare but recurrent variants contributed to the AD-association of *TREM2*, *ABCA7*, *ATP8B4* and *RIN3* (**Figure 2D**). The effect-sizes of rare, coding variant-burdens were large compared to the effect-sizes of the GWAS sentinel SNPs (**Table S5, S6**). Up to 18% EOAD and 14% LOAD cases carried at least one predicted damaging variant in one of the 10 genes, compared to 9% of the controls (**Table S10**). The fractions of EOAD cases in our sample that could be attributed to a rare variant in a specific gene ranged between 0.1% to 2.4%: (~2%: *SORL1, TREM2, ABCA7*; ~1%: *ATP8B4*, *ABCA1*, *RIN3*; and <0.5% for the remaining genes), and for LOAD cases this ranged between 0.0% to 1.3% (**Table 3**).

We performed an age-matched sensitivity analysis to investigate possible effects from other age-related conditions, which supported a role in AD for all 10 identified genes (**Figure S2**). Since *APOE* status was used as selection criterion in several contributing datasets, burden tests were not adjusted for *APOE-ε4* dosage; in a separate analysis we observed no interaction-effects between the rare-variant AD-association and APOE-ε4 dosage (**Table S11, Online Methods**). Also, the rare-variant burden-association was not confounded by somatic mutations due to age-related clonal hematopoiesis (**Table S12**).

Together, we report *ATP8B4* and *ABCA1* as novel AD risk factors with exome-wide significance and we report suggestive evidence for the association of rare variants in the *ADAM10* gene with AD risk. Furthermore, we identified *RIN3*, *CLU*, *ZCWPW1* and *ACE* as potential drivers in GWAS loci, illustrating how analyses of rare protein-modifying variants can solve this drawback of GWAS studies[20]. Larger datasets will be required to further confirm these signals. Given the association of LOF variants with increased AD-risk, we suggest that the GWAS risk alleles in the respective loci might also be associated with reduced activity of the gene, which will have to be evaluated in further experiments. We observed an increased burden of rare damaging genetic variants in individuals with an earlier age at onset. Nevertheless, damaging variants (including APOE-ε4/ε4) were observed in only 30% of the EOAD cases (**Table S10**), suggesting that additional damaging variants remain to be discovered (**Figure 1B**). Further,

the effect of structural variants such as CNVs and repetitive sequences will need to be investigated in future analyses.

The associated genes strengthen our current understanding of AD pathophysiology. When treatment options become available in the future, identification of damaging variants in these genes will be of interest to clinical practice.

Online methods

In-depth descriptions of all methods are described in the Supplemental Methods.

Sample processing, genotype calling and quality control (QC)

We collected the exome, WGS or exome-extract sequencing data of a total of 52,361 individuals, brought together by the Alzheimer Disease European Sequencing consortium (ADES), the Alzheimer's Disease Sequencing Project (ADSP)[21] and several independent study-cohorts (**Table S1)**. Exome-extract samples only contained the raw reads that cover the 10 genes identified in Stage-1. Across all cohorts, AD cases were defined according to NIAA criteria[22] for possible or probable AD or according to NINCDS-ADRDA criteria[23] depending on the date of diagnosis. When possible, supportive evidence for an AD pathophysiological process was sought (including CSF biomarkers) or the diagnosis was confirmed by neuropathological examination (**Table S1**). AD cases were annotated with the age at onset or age at diagnosis (2014 samples), otherwise, samples were classified as late onset AD (366 samples). Controls were not diagnosed with AD. All contributing datasets were sequenced using a paired-end Illumina platform, different exome capture kits were used, and a subset of the sample was sequenced using whole genome sequencing (**Table S2**).

A uniform pipeline was used to process both the Stage-1 and Stage-2 datasets. Raw sequencing data from all studies were processed relative to the GRCh37 reference genome, read alignments of possible

chimeric origin were filtered, and a GATK-based pipeline was used to call variants, while correcting for estimated sample contamination percentages. Samples were included in the datasets after they passed a stringent quality control pipeline: samples were removed when they had high missingness, high contamination, a discordant genetic sex annotation, non-European ancestry, high numbers of novel variants (w.r.t. to DBSNP v150), deviating heterozygous/homozygous or transition/transversion ratios. Further, we removed family members up to the 3rd degree, and individuals who carried a pathogenic variant in *PSEN1*, *PSEN2*, *APP* or in other genes causative for Mendelian dementia diseases (Stage-1 only), or when there was clinical information suggestive of non-AD dementia. Variants considered in the analysis also passed a stringent quality control pipeline: multi-allelic variants were split into bi-allelic variants, variants that were in complete linkage and near each other were merged. Further, we removed variants that had indications of an oxo-G artifact, were located in Short Tandem Repeat (STR) and/or Low Copy Repeat (LCR) regions, had a discordant balance between reads covering the reference and alternate allele, had a low depth for alternate alleles, deviated significantly from Hardy-Weinberg equilibrium, were considered false positive based on GATK VQSR, or were estimated to have a batch effect. Variants with >20% genotype missingness (read depth <6) and differential missingness between the EOAD, LOAD and control groups were removed. To account for uncertainties resulting from variable read coverage between samples, we analyzed variants according to genotype posterior likelihoods, i.e., the likelihood for being homozygous for the reference allele, heterozygous or homozygous for the alternate allele. To account for genotype uncertainty, the burden test was performed multiple times with independently sampled genotypes and the average p-value across these tests is reported.

Variant prioritization and thresholds

We selected variants in autosomal protein-coding genes that were part of the Ensembl basic set of protein coding transcripts (Gencode v19/v29[24], see Supplement) and that were annotated by the Variant Effect Predictor (VEP) (version 94.542)[25]. Only protein coding missense and loss-of-function

(LOF) variants were considered (LOF: nonsense, splice acceptor/donor or frameshifts). Missense and LOF variants were required to have respectively a 'moderate' and 'high' VEP impact classification. Then, missense variants were prioritized using REVEL (Rare Exome Variant Ensemble Learner[26], annotation obtained from DBNSFP4.1a [27] and LOF variants were prioritized using LOFTEE (version 1.0.2)[28]. For the analysis we considered only missense variants with a REVEL score ≥ 25 (score range 0 – 100) and LOF variants annotated 'high-confidence' by LOFTEE. Variants were required to have at least one carrier (i.e. at least one one sample with a posterior dosage of >0.5), and a minor allele frequency (MAF) of <1%, both in the considered dataset and the gnomAD v2.1 populations (non-neuro set).

Gene burden testing

The burden analysis was based on four deleteriousness thresholds by incrementally including variants from categories with lower levels of predicted variant deleteriousness: respectively LOF, LOF + REVEL≥75, LOF + REVEL≥50, and LOF + REVEL≥25. This allowed us to identify the variant-threshold providing maximum evidence for a differential burden-signal. To infer any dependable signal for a specific deleterious threshold, a minimum of 10 damaging alleles appertaining to this deleteriousness threshold was required: i.e., a cumulative minor allele count (cMAC) ≥10. Multiple testing correction was performed across all performed tests (up to 4 per gene). Burden testing was implemented using ordinal logistic regression. This enabled the burden testing to particularly weight EOAD cases, as previous findings indicated that high-impact variants are enriched in early onset (EOAD) cases relative to late onset (LOAD) cases[29]. This implies that the burden of high-impact deleterious genetic variants is ordered according to: $burden_{EOAD} > burden_{LOAD} > burden_{control}$. Ordinal logistic regression enabled optimal identification of such signals, while also allowing the detection of EOAD-specific burdens ($burden_{EOAD} > burden_{LOAD} \sim burden_{control}$) and regular case-control signals ($burden_{EOAD} \sim burden_{LOAD} > burden_{control}$). For protective burden signals, the order of the signals is reversed, such that $burden_{EOAD} < burden_{LOAD} < burden_{control}$. We considered an additive model, while correcting for 6 population covariates, estimated after removal of population outliers. P-values were estimated using a likelihood-

ratio test. Genes were selected for confirmation in Stage-2 if the False Discovery Rate for AD association was <0.1 in Stage-1 (Benjamini-Hochberg procedure[30]). For the GWAS targeted analysis, a more stringent threshold was used (FDR<0.05) due to the absence of a separate confirmation stage. For the meta-analysis, genes were considered significantly associated with AD when the corrected P was <0.05 after family-wise correction using the Holm-Bonferroni procedure[31]. Effect-sizes (odds ratios, ORs) of the ordinal logistic regression can be interpreted as weighted averages of the OR of being an AD case versus control, and the OR of being an early-onset AD case or not. To aid interpretation, we additionally estimated 'standard' case/control ORs across all samples, per age category (EOAD versus controls and LOAD versus controls), and for age-at-onset categories ≤65 (EOAD), (65-70], (70-80] and >80 using multinomial logistic regression, while correcting for 6 PCA covariates.

GWAS driver gene identification

For the 75 loci identified in the most recent GWAS[3], genes were selected for burden testing based on earlier published gene prioritizations. First, gene prioritizations were obtained from Schwarzentruber et al.[32] for 33 known loci. For 28 remaining loci, we obtained the Tier 1 prioritization from Bellenguez et al. [3], and for the loci without prioritization candidates (14 loci), we selected the nearest gene. In total, 81 protein-coding genes were selected (**Table S5**), of which 67 genes had sufficient damaging allele carriers to be tested for at least one variant selection threshold. Gene burden testing was performed as described above, and multiple testing correction to identify potential driver genes was performed using the Benjamini-Hochberg procedure, with a cutoff of 5%.

Validation of variant selection

We validated the REVEL variant impact prediction for missense and the LOFTEE impact prediction for LOF variants, for all variants with a MAF < 1%, for which there were at least 15 damaging allele carriers. For protein-modifying variants that were not in the most significant burden selection of a gene due to

a low predicted impact, we investigated whether they, nevertheless, showed a significant AD-association (based on a case/control analysis using logistic regression). Vice versa, for variants that were in the burden selection, we investigated whether their effect-size was significantly reduced or oppositely directed from other missense or LOF variants in the burden selection (Fisher exact test). Individual variant-effects were analyzed in the Stage-1 dataset, followed by a confirmation analysis in the Stage-2 dataset. Multiple testing correction was performed per gene, with FDR<0.1 used as threshold for Stage-1 and Holm-Bonferroni (P<0.05) for Stage-2.

Descriptive measures

A variant-carrier was defined as an individual for whom the summed dosage of *all* the variants in the considered variant category is ≥0.5 (Supplementary methods). Carrier frequencies (CFs) were determined as #carriers / #total samples. Attributable fraction for cases in an age group was estimated as the probability of a case with an age-at-onset in age window *i* being exposed to a specific gene gene burden$\left(CF_{case,gene,i}\right)$, multiplied by an estimate of the attributable fraction among the exposed for these cases: $\left(\frac{OR_{gene,i}-1}{OR_{gene,i}}\right)$ (with the odds ratio being an approximation of the relative risk)[33,34]. For large effect-sizes, this estimate approaches the difference in carrier frequency between cases and controls: $\left(CF_{case,gene,i}\right) - \left(CF_{control,gene}\right)$.

Sensitivity analyses

We determined if observed effects could be explained by age differences between cases and controls. We constructed an age-matched sample, dividing samples in strata based on age/age-at-onset, with each stratum covering 2.5 years. Case/control ratios in all strata were kept between 0.1 and 10 by down sampling respectively controls or cases. Subsequently, samples were weighted using the propensity weighting within strata method proposed by Posner and Ash[35]. Finally, a case-control logistic regression was performed both on the unweighted and weighted case-control labels, and

estimated odds ratios and confidence intervals were compared (**Figure S2**). Also, we determined if somatic mutations due to age-related clonal hematopoiesis could have confounded the results. We calculated for all heterozygous calls in the burden selection the balance between reference and alternate reads, and compared these to reference values (**Table S12**). While APOE was not included as a confounder, we performed a separate APOE interaction analysis (**Table S11**), through a likelihood ratio test between a model: $\text{label} \sim \text{gene_burden_score} + \text{apoe_e4_dosage}$ and an interaction model: $\text{label} \sim \text{gene_burden_score} + \text{apoe_e4_dosage} + \text{apoe_e4_dosage} * \text{gene_burden_score}$. This test was performed on a reduced dataset, from which datasets in which APOE status was used as selection criterion were removed.

Power analysis

Power calculations were performed for ordinal logistic regression and Firth logistic regression (case-control and EOAD vs. rest), **Figure 1B** and **Table S4.** Given odds ratios for EOAD and LOAD cases, and the cumulative minor allele count per gene, we sampled the number of alleles in EOAD cases, LOAD cases and controls according to a multinomial distribution. We randomized these allele-carriers across the dataset, and performed the burden test (as described above). Power for genes with cMAC <10 was set to 0, as these genes were not analyzed.

Data availability Statement

The genetic variants analyzed in this study will be included in the Supplementary data files (Tables SG1-SG10) upon the formal publication of this manuscript. Summary statistics of the Stage-1 analysis will be made available through [**holstegelab.eu/data**]. The full datasets generated during and/or analyzed during the current study are not publicly available due to privacy restrictions applicable to genetic data from human subjects.

Code availability statement

All software and algorithms used in the analysis are described in the Supplement attached to this Letter. Self-contained code can be found at **holstegelab.eu/tools**.

References

1. Georges, J.M., O. Bintener C. *Estimating the prevalence of dementia in Europe*, (Luxembourg, 2020).
2. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry* **63**, 168-74 (2006).
3. Bellenguez, C. *et al.* New insights on the genetic etiology of Alzheimer's and related dementia. (2020).
4. Holstege, H. *et al.* Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. *Eur J Hum Genet* **25**, 973-981 (2017).
5. Nicolas, G. *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. *Mol Psychiatry* (2015).
6. Cuyvers, E. *et al.* Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. *Lancet Neurol* **14**, 814-22 (2015).
7. Jonsson, T. *et al.* Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med* **368**, 107-16 (2013).
8. Guerreiro, R. *et al.* TREM2 variants in Alzheimer's disease. *N Engl J Med* **368**, 117-27 (2013).
9. Gao, L. *et al.* Identification of Rare Variants inATP8B4as a Risk Factor for Systemic Sclerosis by Whole-Exome Sequencing. *Arthritis & Rheumatology* **68**, 191-200 (2016).
10. Wahrle, S.E. *et al.* Overexpression of ABCA1 reduces amyloid deposition in the PDAPP mouse model of Alzheimer disease. *Journal of Clinical Investigation* (2008).
11. Koldamova, R., Staufenbiel, M. & Lefterov, I. Lack of ABCA1 Considerably Decreases Brain ApoE Level and Increases Amyloid Deposition in APP23 Mice. *Journal of Biological Chemistry* **280**, 43224-43235 (2005).
12. Nordestgaard, L.T., Tybjaerg-Hansen, A., Nordestgaard, B.G. & Frikke-Schmidt, R. Loss-of-function mutation inABCA1and risk of Alzheimer's disease and cerebrovascular disease. *Alzheimer's & Dementia* **11**, 1430-1438 (2015).
13. Saftig, P. & Lichtenthaler, S.F. The alpha secretase ADAM10: A metalloprotease with multiple functions in the brain. *Progress in Neurobiology* **135**, 1-20 (2015).
14. Kim, M. *et al.* Potential late-onset Alzheimer's disease-associated mutations in the ADAM10 gene attenuate α-secretase activity. *Human Molecular Genetics* **18**, 3987-3996 (2009).
15. Agüero, P. *et al.* α-Secretase nonsense mutation (ADAM10 Tyr167*) in familial Alzheimer's disease. *Alzheimer's Research & Therapy* **12**(2020).
16. Shen, R. *et al.* Upregulation of RIN3 induces endosomal dysfunction in Alzheimer's disease. *Translational Neurodegeneration* **9**(2020).
17. Shen, R. & Wu, C. RIN3 Binds to BIN1 and CD2AP to increase APP-CTFS in early endosomes. *Alzheimer's & Dementia* **16**(2020).

18. Foster, E.M., Dangla-Valls, A., Lovestone, S., Ribe, E.M. & Buckley, N.J. Clusterin in Alzheimer's Disease: Mechanisms, Genetics, and Lessons From Other Pathologies. *Frontiers in Neuroscience* **13**(2019).
19. Hu, J., Igarashi, A., Kamata, M. & Nakagawa, H. Angiotensin-converting Enzyme Degrades Alzheimer Amyloid β-Peptide (Aβ); Retards Aβ Aggregation, Deposition, Fibril Formation; and Inhibits Cytotoxicity. *Journal of Biological Chemistry* **276**, 47863-47868 (2001).
20. Backman, J.D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628-634 (2021).
21. Bis, J.C. *et al.* Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Molecular Psychiatry* **25**, 1859-1875 (2018).
22. McKhann, G.M. *et al.* The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* **7**, 263-269 (2011).
23. McKhann, G. *et al.* Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-939 (1984).
24. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766-D773 (2019).
25. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
26. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99**, 877-885 (2016).
27. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12**(2020).
28. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
29. Bellenguez, C. *et al.* Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiol Aging* **59**, 220 e1-220 e9 (2017).
30. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289-300 (1995).
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979).
32. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nature Genetics* **53**, 392-402 (2021).
33. Cole, P. & MacMahon, B. Attributable risk percent in case-control studies. *Journal of Epidemiology & Community Health* **25**, 242-244 (1971).
34. LaMorte, W.W. Measuring Association in Case-Control Studies. in *Measures of Association* (Boston University School of Public Health, 2018).
35. Posner, M. & Ash, A. Comparing weighting methods in propensity score analysis. http://www.stat.columbia.edu/~gelman/stuff_for_blog/posner.pdf. *Unpublished working paper, Columbia University* (2012).

Tables

Table 1: Quality Control Steps

Table 1A: Sample Quality Control

Sample	**nr. of samples**		
Samples processed	**Stage-1**	**Stage-2**	**Mega**
Samples included in study	25,982	26,379	52,361
Samples retained after QC (**Table S3A**)	21,345	11,213	31,905
Sample totals (of which exome-extracts of targeted genes)			
EOAD	4,060	1,627 (446)	5,643 (446)
LOAD	8,592	1,757 (385)	10,165 (385)
Controls	8,693	7,829 (1,347)	16,097 (1,347)

Table 1B: Variant Quality Control

			Targeted genes	
Variant QC		**Stage-1**	**Stage-2**	**Mega**
Variants called (bi-allelic)		12,938,556	7,803	14,531
Variants retained after QC (**Table S3B-C**)		7,543,193	5,072	8,963
Variant selection				
1. In protein coding autosomal genes (Gencode V19/V29)		6,883,630	5,072	8,963
Missense variants	2. Missense variants	1,486,559	894	1,873
	3. REVEL > 25	540,934	591	1,263
	4. MAF < 1% / dosage > 0.5	530,072	567	1,228
	5. Missingness (< 20% + no differential missingness)	353,913	428	943
LOF variants	2. Loss-of-function variants (nonsense, frameshift, splice acceptor/donor)	144,429	106	255
	3. Loftee HC + VEP high impact	109,550	97	236
	4. MAF < 1% / dosage > 0.5	108,016	96	234
	5. Missingness (< 20% + no differential missingness)	57,543	64	168
Variant Categories	REVEL 25-50	198,068	168	360
	REVEL 75-100	99,910	140	320
	REVEL 50-75	54,212	119	258
	LOF	57,543	64	168
Variant Thresholds	LOF+REVEL≥25	409,733	491	1,106
	LOF+REVEL≥50	211,665	323	746
	LOF+REVEL≥75	111,755	183	426
	LOF	57,543	64	168

A. Sample QC: Samples were primarily excluded due to non-European ancestry or close family relations, for details see **Table S3A**. Exome-extract samples (between parenthesis) only contain reads that cover the 10 genes discovered in Stage-1. In Stage-2, samples were removed that were duplicated w.r.t. Stage-1. In the mega-analysis, a merged sample QC removed all family relations to the third

degree between Stage-1 and Stage-2 (i.e. the size of the mega-sample does not equal the sum of the Stage-1 and Stage-2 samples). **B. Variant QC:** Stage-1 consists of all variants in the union of the exome capture kits. The targeted Stage-2 and Mega analysis regards only the 10 genes identified in Stage-1 . Variant QC from the non-targeted Stage-2 and mega-analysis can be found in **Table S3B**. For each gene, we considered in our variant selection 4 different selection thresholds. MAF: minor allele frequency. See supplement for in-depth QC methods.

Table 2. Stage-1, Stage-2 Association statistics

Table 2A. Burden tests Stage-1, Stage-2 and Meta-analysis.												
		Stage-1				Stage-2			meta			
gene	Variant damagingness threshold	P value	FDR	#variants / #carriers	case / control	P value#	#variants / #carriers	case / control OR (95% CI)	P value	h-bonf	case / control OR (95% CI)	pvalue heterog.
SORL1	LOF+REVEL≥25	4.8E-06	**0.017**	242 / 917	1.3 (1.1-1.5)	**1.3E-06**	122 / 478	1.5 (1.2-1.9)	1.5E-10	**4.7E-06**	1.4 (1.2-1.5)	1.6E-01
	LOF+REVEL≥50	4.0E-18	**<<0.0001**	167 / 290	2.6 (2.0-3.2)	**1.4E-09**	79 / 137	2.4 (1.7-3.5)	8.1E-26	**2.5E-21**	2.5 (2.1-3.1)	9.8E-01
	LOF+REVEL≥75	1.1E-14	**<<0.0001**	96 / 164	3.3 (2.4-4.6)	**5.2E-10**	45 / 82	3.9 (2.3-6.6)	1.1E-22	**3.4E-18**	3.5 (2.7-4.6)	4.3E-01
	LOF	4.7E-15	**<<0.0001**	37 / 48	15.6 (3.7-37.3)	**1.6E-06**	16 / 20	16.3 (3.8-35.0)	3.3E-18	**1.0E-13**	16.0 (9.5-27.0)	9.4E-01
TREM2	LOF+REVEL≥25	2.6E-16	**<<0.0001**	17 / 291	3.6 (2.9-4.6)	**1.6E-07**	12 / 155	2.4 (1.6-3.4)	5.2E-22	**1.6E-17**	3.2 (2.6-3.9)	6.5E-01
ABCA7	LOF+REVEL≥25	9.5E-08	**0.001**	265 / 959	1.4 (1.2-1.6)	**9.8E-08**	170 / 502	1.6 (1.3-2.0)	4.1E-13	**1.3E-08**	1.4 (1.3-1.6)	6.5E-02
	LOF+REVEL≥75	4.6E-06	**0.017**	93 / 297	1.6 (1.3-2.1)	**4.8E-04**	54 / 167	1.8 (1.3-2.6)	7.3E-09	**2.3E-04**	1.7 (1.4-2.1)	9.1E-01
ATP8B4	LOF+REVEL≥25	7.2E-06	**0.02**	72 / 575	1.5 (1.3-1.8)	**3.3E-03**	40 / 286	1.4 (1.0-1.8)	9.6E-09	**3.0E-04**	1.5 (1.3-1.7)	9.7E-01
	LOF+REVEL≥50	2.8E-05	**0.068**	61 / 521	1.5 (1.3-1.9)	**1.6E-02**	34 / 265	1.3 (1.0-1.7)	2.8E-06	**8.7E-02**	1.5 (1.3-1.7)	6.6E-01
	LOF+REVEL≥75	3.2E-06	**0.014**	38 / 490	1.7 (1.4-2.0)	**2.4E-02**	22 / 243	1.3 (1.0-1.8)	5.7E-07	**1.8E-02**	1.5 (1.3-1.8)	4.2E-01
ABCA1	LOF+REVEL≥75	6.1E-06	**0.019**	93 / 280	1.7 (1.3-2.2)	**6.6E-03**	48 / 159	1.6 (1.1-2.3)	2.6E-07	**8.0E-03**	1.7 (1.4-2.1)	6.3E-01
ADAM10	LOF+REVEL≥50	2.0E-05	**0.051**	15 / 17	3.2 (1.3-8.1)	**4.0E-02**	4 / 4	8.1 (0.6-42.6)	2.8E-05	8.7E-01	3.6 (1.5-8.5)	5.5E-01
	LOF+REVEL≥75	2.7E-06	**0.014**	11 / 12	7.5 (1.4-46.8)	1.5E-01	3 / 3	5.6 (0.3-41.8)	4.4E-04	1.0E+00	7.1 (2.6-19.3)	1.1E-01

Table 2B. GWAS-targeted analysis on mega-dataset without exome-extracts										
		Burden test (variant MAF <1%)					Burden test (variant MAF <0.1%)			
Locus *sentinal GWAS SNP*	gene	Variant damagingness threshold	P value	FDR	#variant / #carriers	case / control OR (95% CI)	pvalue	#variant / #carriers	fraction very rare	case / control OR (95% CI)
SORL1, TREM2, ABCA7* (see **Table 2A and S6)										
SLC24A4/RIN3 *rs7401792* *rs12590654*	*RIN3*	LOF+REVEL≥25	1.6E-05	0.0003	44 / 622	1.4 (1.2-1.6)	3.4E-02	42 / 129	21%	1.4 (1.0-2.1)
		LOF+REVEL≥50	1.0E-05	0.0002	23 / 583	1.4 (1.2-1.7)	1.5E-02	21 / 89	15%	1.8 (1.2-2.8)
ADAM10, ABCA1* (see **Table 2A and S6)										
PTK2B/CLU *rs73223431* *rs11787077*	*CLU*	LOF+REVEL≥25	5.0E-04	0.005	24 / 26	3.6 (1.6-8.3)	5.0E-04	24 / 26	100%	3.6 (1.6-8.3)
		LOF+REVEL≥50	1.1E-03	0.001	14 / 15	5.4 (1.6-28.6)	1.1E-03	14 / 15	100%	5.3 (1.6-28.6)
		LOF+REVEL≥75	5.0E-04	0.005	12 / 12	9.9 (1.6-44.0)	5.0E-04	12 / 12	100%	9.8 (1.6-44.0)
		LOF	2.6E-03	0.02	10 / 10	7.3 (1.9-27.2)	2.6E-03	10 / 10	100%	7.3 (1.9-27.2)
SPDYE3 *rs7384878*	*ZCWPW1*	LOF+REVEL≥25	6.1E-03	0.042	22 / 77	1.8 (1.2-2.9)	5.0E-03	21 / 76	99%	1.8 (1.2-2.9)
		LOF+REVEL≥50	3.1E-03	0.022	16 / 70	1.9 (1.2-3.1)	3.1E-03	16 / 70	100%	1.9 (1.2-3.1)
		LOF+REVEL≥75	1.1E-03	0.001	11 / 15	5.0 (1.9-13.5)	1.1E-03	11 / 15	100%	5.0 (1.9-13.5)
		LOF	7.8E-04	0.008	11 / 15	5.0 (1.9-13.5)	7.8E-04	11 / 15	100%	5.0 (1.9-13.5)
ACE *rs4277405*	*ACE*	LOF+REVEL≥75	9.0E-04	0.008	38 / 99	2.0 (1.3-2.9)	9.3E-04	38 / 99	100%	2.0 (1.3-2.9)

A. Stage-1, Stage-2 and Meta-analysis. Burden tests that were significant in Stage-1 after multiple testing correction using the Benjamini-Hochberg False Discovery Rate (FDR) (<0.1) over 31,204 tests/variant categories. #In Stage-2, we considered only the direction of the AD-association observed in Stage-1. The meta-analysis indicates the combined significance from Stages 1 and 2 (data was combined using the fixed-effect inverse variance method); multiple testing correction for the meta-analysis was performed across all 31,204 tests using the Holm-Bonferroni correction (<0.05). Bold black text: significant p-values; **B. GWAS-targeted analysis in mega-dataset without exome extracts.** Genes in all loci were prioritized as described in the Online Methods, (**Table S5**). *These genes also included the

SORL1, *TREM2*, *ABCA7*, *ADAM10*, and *ABCA1* genes, which were also identified in the rare variant burden analysis (shown in A) and therefore not shown here. These genes are listed in the context of this analysis in **Table S6**. For both tables the p values were determined using ordinal logistic regression, and a case/control OR was computed for reference. Grey text: result burden test MAF<0.1% unchanged compared to burden test MAF<1%.

Table 3. **Mega Analysis; effect-sizes and p values.**

	Mega-analysis		Carrier frequency		Odds ratio (95% CI)			median age at onset (IQR)	Attr. fraction
	gene	group	#variant / #carriers	EOAD / LOAD / controls	case / control	EOAD / control	LOAD / control		EOAD / LOAD
Table 3a: primary analysis	***SORL1***	LOF+REVEL≥50	212 / 418	2.75% / 1.51% / 0.68%	2.5 (2.0-3.0)	3.3 (2.6-4.1)	2.0 (1.6-2.5)	65 (59-73)	1.91% / 0.75%
		- REVEL 50-100	161 / 354	2.02% / 1.31% / 0.66%	2.1 (1.7-2.5)	2.5 (2.0-3.2)	1.8 (1.4-2.3)	67 (59-74)	1.22% / 0.58%
		- LOF	51 / 68	0.78% / 0.21% / 0.02%	19.8 (11.9-32.7)	40.7 (12.5-133)	11.3 (3.3-38.3)	62 (56-69)	0.76% / 0.19%
	TREM2	LOF+REVEL≥25	26 / 441	2.27% / 1.90% / 0.75%	2.8 (2.3-3.5)	3.3 (2.6-4.3)	2.6 (2.1-3.3)	69 (62-75)	1.58% / 1.17%
		LOF+REVEL≥25 [refined]	25 / 404	2.22% / 1.77% / 0.62%	3.1 (2.6-3.8)	3.8 (2.9-4.9)	2.8 (2.2-3.6)	68 (62-75)	1.63% / 1.15%
		- REVEL 25-100	14 / 377	2.06% / 1.63% / 0.59%	3.0 (2.5-3.8)	3.7 (2.8-4.9)	2.7 (2.1-3.6)	68 (62-75)	1.50% / 1.04%
		- LOF	12 / 66	0.21% / 0.29% / 0.16%	2.1 (1.2-3.4)	1.7 (0.8-3.5)	2.2 (1.3-3.9)	71 (63-76)	0.09% / 0.16%
		- LOF [refined]	11 / 29	0.16% / 0.16% / 0.02%	5.6 (2.6-12.1)	5.8 (1.7-19)	5.4 (1.8-16.8)	71 (63-74)	0.13% / 0.13%
	ABCA7	LOF+REVEL≥25	351 / 1489	6.18% / 5.04% / 3.90%	1.4 (1.3-1.6)	1.6 (1.4-1.9)	1.3 (1.2-1.5)	69 (61-78)	2.40% / 1.29%
		- REVEL 25-100	302 / 1372	5.58% / 4.65% / 3.63%	1.4 (1.3-1.6)	1.6 (1.4-1.8)	1.3 (1.2-1.5)	69 (62-78)	2.06% / 1.18%
		- LOF	49 / 119	0.62% / 0.39% / 0.27%	1.7 (1.1-2.4)	2.2 (1.4-3.5)	1.4 (0.9-2.1)	67 (57-74)	0.34% / 0.11%
	ATP8B4	LOF+REVEL≥25	94 / 850	3.56% / 3.08% / 2.09%	1.4 (1.2-1.6)	1.5 (1.3-1.8)	1.4 (1.2-1.6)	70 (61-78)	1.24% / 0.84%
		- REVEL 25-100	74 / 797	3.35% / 2.93% / 1.93%	1.5 (1.3-1.7)	1.6 (1.3-1.9)	1.4 (1.2-1.7)	70 (62-78)	1.20% / 0.84%
		- LOF	20 / 54	0.21% / 0.16% / 0.16%	1.1 (0.6-1.9)	1.2 (0.6-2.4)	1.0 (0.5-1.8)	70 (59-78)	0.03% / -0.01%
	ABCA1	LOF+REVEL≥75	122 / 442	1.91% / 1.50% / 1.13%	1.6 (1.3-2.0)	1.9 (1.5-2.5)	1.5 (1.2-1.9)	70 (60-76)	0.91% / 0.48%
		LOF+REVEL≥75 [refined]	120 / 282	1.52% / 1.10% / 0.52%	2.4 (1.9-3.1)	2.9 (2.2-4.0)	2.2 (1.6-2.9)	70 (59-76)	1.01% / 0.60%
		- REVEL 75-100	95 / 395	1.63% / 1.32% / 1.05%	1.5 (1.2-1.8)	1.7 (1.3-2.2)	1.4 (1.1-1.8)	70 (61-76)	0.68% / 0.37%
		- REVEL 75-100 [refined]	93 / 235	1.24% / 0.92% / 0.44%	2.3 (1.7-3.0)	2.7 (1.9-3.8)	2.1 (1.5-2.8)	70 (59-76)	0.78% / 0.48%
		- LOF	27 / 47	0.28% / 0.18% / 0.08%	3.5 (1.9-6.4)	4.7 (2.2-10.3)	2.8 (1.3-6.1)	67 (59-77)	0.22% / 0.11%
	ADAM10	LOF+REVEL≥50	19 / 22	0.23% / 0.05% / 0.02%	4.7 (2.0-10.8)	9.0 (2.9-28)	2.2 (0.5-8.2)	63 (60-68)	0.20% / 0.03%
Table 3b: GWAS-targeted analysis	*RIN3*	LOF+REVEL≥50	23 / 583	2.67% / 2.10% / 1.62%	1.4 (1.2-1.7)	1.6 (1.3-2.0)	1.3 (1.1-1.6)	70 (59-79)	1.04% / 0.46%
		- REVEL 50-100	17 / 577	2.62% / 2.08% / 1.61%	1.4 (1.2-1.7)	1.6 (1.3-2.0)	1.3 (1.1-1.6)	70 (59-79)	1.01% / 0.45%
		- LOF	6 / 8	0.06% / 0.03% / 0.01%	2.1 (0.5-9.3)	2.9 (0.5-18.0)	1.7 (0.3-10.3)	69 (57-86)	0.04% / 0.01%
	CLU	LOF+REVEL≥25	24 / 26	0.23% / 0.09% / 0.03%	3.6 (1.6-8.3)	5.8 (2.0-17.1)	2.5 (0.8-7.6)	63 (58-73)	0.19% / 0.05%
		- REVEL 25-100	14 / 16	0.12% / 0.06% / 0.03%	2.6 (0.9-7.5)	3.6 (0.9-13.6)	2.1 (0.6-8.0)	68 (58-76)	0.08% / 0.03%
		- LOF	10 / 10	0.12% / 0.03% / 0.01%	7.3 (1.9-27.2)	14.2 (2.9-470.4)	3.8 (0.6-122.4)	63 (59-68)	0.11% / 0.02%
	ZCWPW1	LOF	11 / 15	0.15% / 0.05% / 0.01%	5.0 (1.9-13.5)	9.1 (2.0-42.0)	2.9 (0.8-14.7)	63 (58-81)	0.14% / 0.03%
	ACE	LOF+REVEL≥75	38 / 99	0.60% / 0.39% / 0.20%	2.0 (1.3-2.9)	2.4 (1.5-4.1)	1.7 (1.0-2.7)	67 (60-75)	0.35% / 0.16%
		- REVEL 75-100	10 / 49	0.33% / 0.22% / 0.07%	3.2 (1.7-5.7)	3.9 (1.8-8.8)	2.7 (1.3-5.9)	66 (61-72)	0.24% / 0.14%
		- LOF	28 / 50	0.27% / 0.16% / 0.14%	1.4 (0.8-2.4)	1.7 (0.9-3.4)	1.2 (0.6-2.2)	70 (55-76)	0.11% / 0.02%

Per gene, the characteristics are shown for the variant deleteriousness threshold with the most evidence for AD association. For genes with sufficient carriers, signals are shown for LOF and missense variants separately. Variants contributing to the burden were validated in a multi-stage analysis (**Table S12**, Online Methods), which resulted in the construction of a *refined* burden for *TREM2* (1 variant removed) and *ABCA1* (2 variants removed). The attributable fraction of a gene is an estimate of the fraction of EOAD and LOAD cases in this sample that have become part of this dataset due to carrying a rare damaging variant in the respective gene (Online methods). Note that several variants were excluded from this analysis (i.e., due to differential missingness) that would otherwise have been included in the burden. See **section 1.12 of the supplement** for a gene-specific discussion of the variants that contribute to the association with AD, and gene-specific **Tables SG1-SG10** for the list of variants considered in the burden-analysis. P values for the mega-analysis are shown in **Table S14**.

Figures

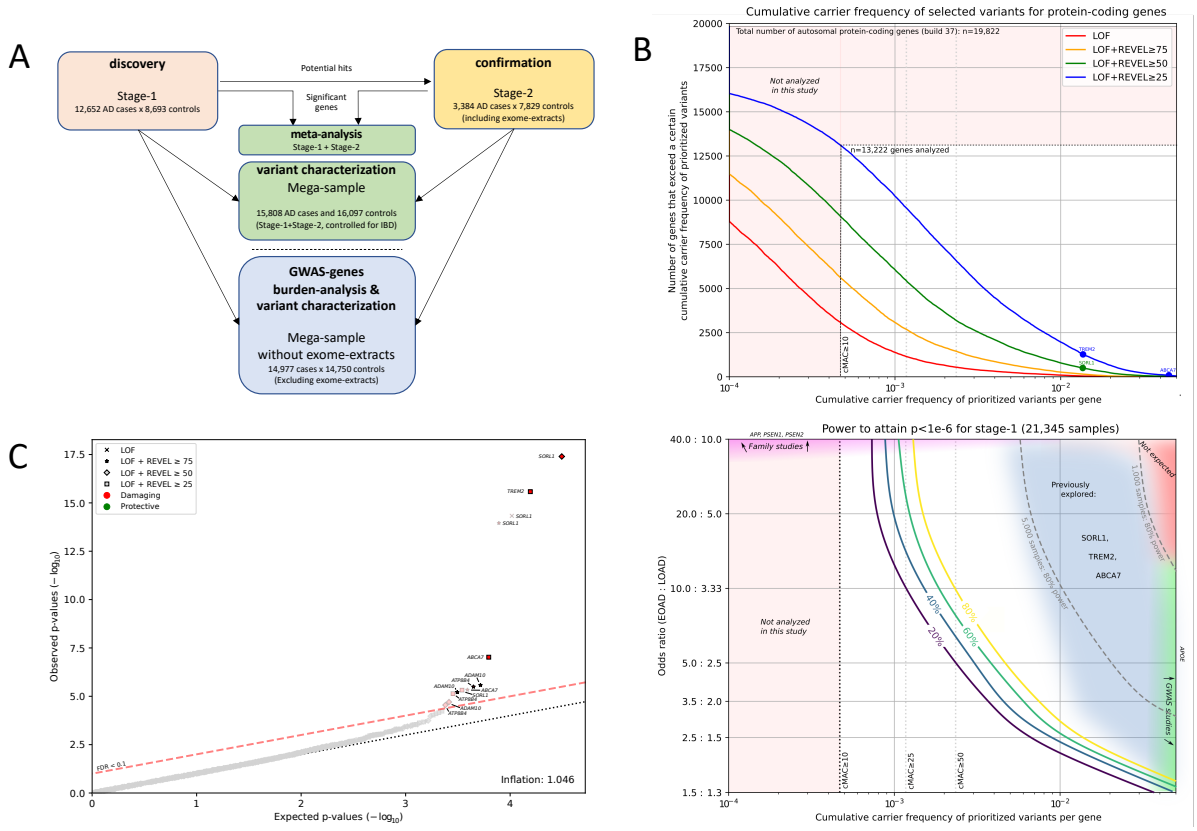
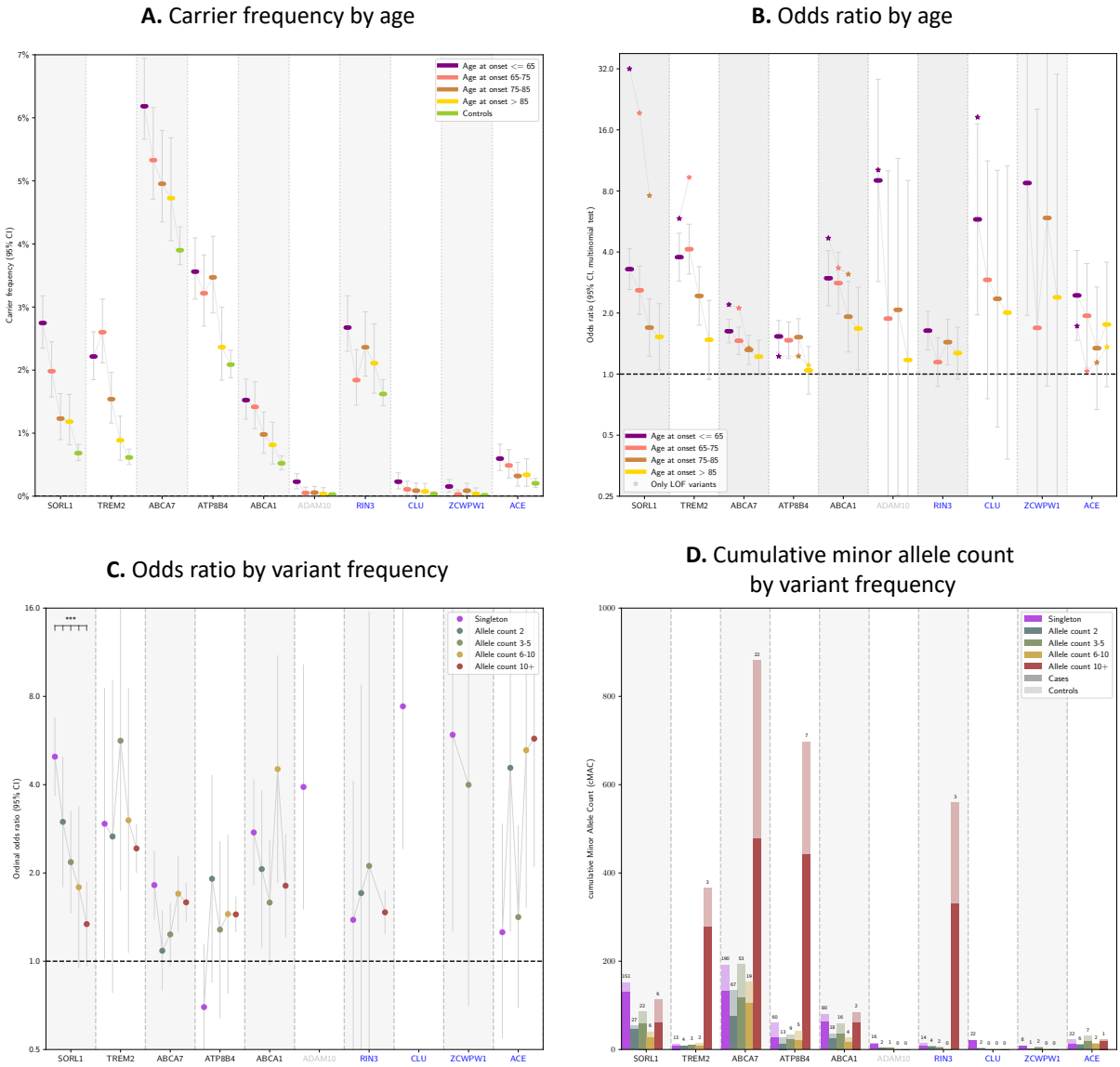


Fig 1. Study set-up and power A: Schematic of study set-up; The AD-association of genes identified in Stage-1 was confirmed in Stage-2, and combined significance was determined in a meta-analysis. To accurately estimate variant effect-sizes for variant-categories/age at onset bins, even those with very low variant counts for which the normal-distribution approximation fails, we analyzed variant characteristics in a mega-sample instead of through meta-analysis. This mega-sample was also used for the GWAS gene burden analysis (without exome-extracts). **B.** Top: the number of genes (Y-axis) that have at least a certain cumulative carrier frequency of prioritized variants (X-axis). Variants were prioritized according to different deleteriousness thresholds. White box: only genes with a cMAC≥10 (cumulative minor allele count of at least 10 alleles across the sample) in the Stage-1 sample of 12,652 cases and 8,693 controls were considered to have a high enough carrier frequency allowing burden analysis. The previously identified *SORL1*, *TREM2*, and *ABCA7* genes are indicated, revealing that carriers of damaging variants in these genes are relatively common, which has aided their identification in the past. **B.** Bottom: Power analysis for Stage-1, to attain a $p<1e^{-6}$, on the same scale as the top

figure B. For comparison purposes, we also plot 80% power thresholds for 5,000 and 1,000 samples (subsampled from Stage-1). Approximate regions are indicated for variants identified with GWAS (red) or family studies (purple), as well as the region in which variant-burdens in *SORL1*, *TREM2* and *ABCA7* were identified by previous sequencing studies[4-8] (dark-blue). Common variants with very high effect-sizes (red) are not expected to exist. Genes with cMAC<10 were not analyzed (light-red). Power calculations show that, by aggregating more cases and controls, one might be able to identify burdens of rare variants with either (i) a large effect but with an extremely low frequency of carriers or (ii) with a modest average effect but a higher number of carriers. **C.** P-value Q-Q plot of Stage-1 discovery-analysis. Gene-names were indicated in grey when the deleteriousness threshold was not the most significant burden test in that gene.

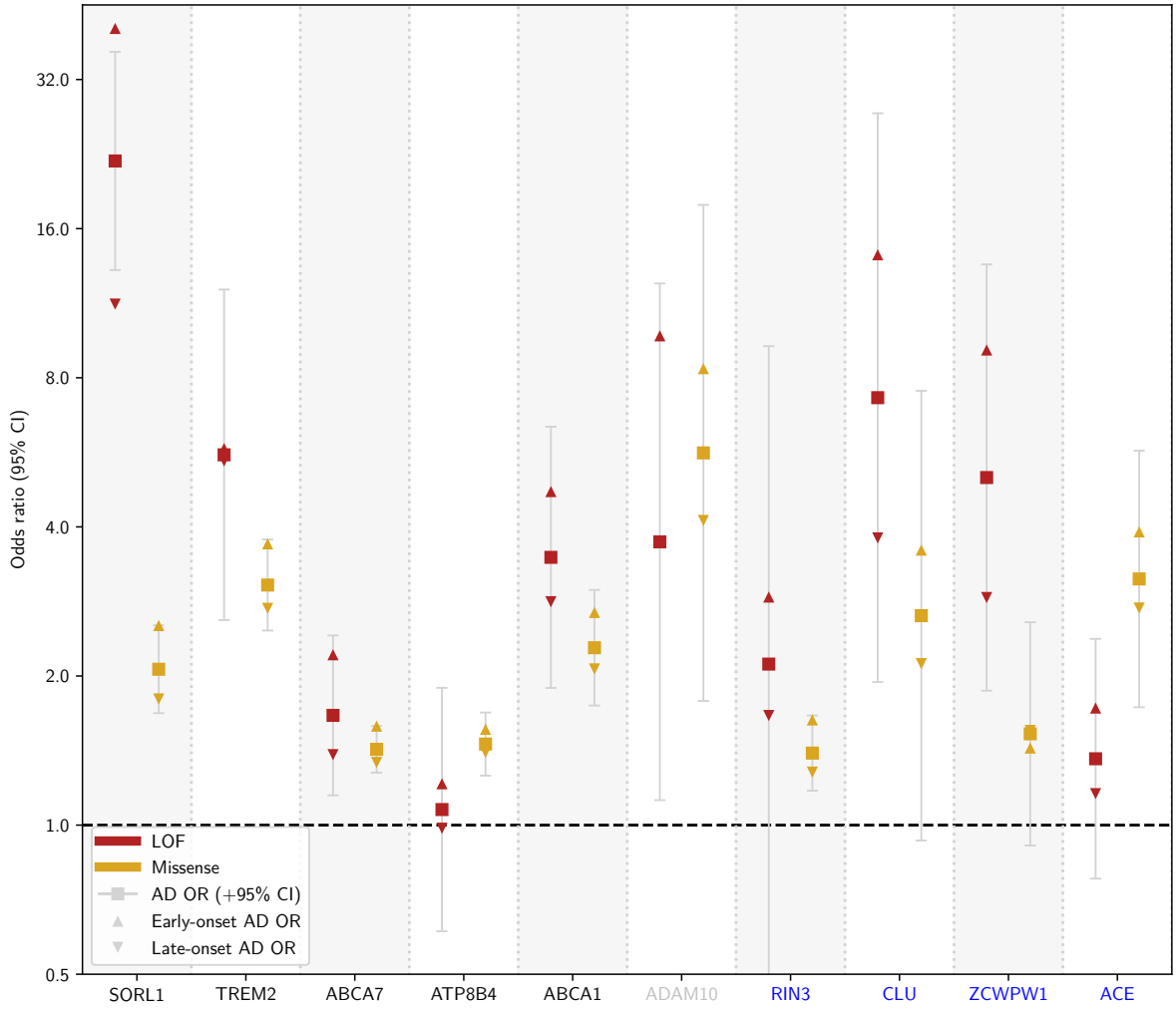
Fig. 2: Characterization of gene specific variant features based on the mega-sample.



We considered the deleteriousness threshold that provides the most evidence for AD-association (**Table 3,** refined burden). **A. Carrier frequency by age at onset.** A carrier is an individual who carries at least one damaging variant in the considered gene. **B. Odds ratios by age at onset.** The effect-size significantly decreased with age at onset for *SORL1, TREM2, ABCA7, ABCA1, ADAM10* (after multiple testing correction, **Table S7**). **C. Odds ratios by variant frequency.** The rareness of variants in *SORL1* significantly associated with the effect size (**Table S9**). **D. Cumulative minor allele count by variant frequency**: the total number of cases (dark) and controls (light) that carry gene-variants with allele

frequencies as observed in the mega-sample. Numbers above the bars indicate the number of contributing variants. Whiskers: 95% CI.

Fig 3. Summary Figure: Odds ratio by age at onset and variant pathogenicity.



Odds ratio's for LOF (red) and missense (yellow) variants. Case/control OR (square, 95% CI), EOAD OR (triangle pointing upwards), LOAD OR (triangle pointing downwards). Missense variants in the considered gene appertained to the variant-deleteriousness threshold that provides the most evidence for its AD-association (**Table 3, refined**). The LOF burden effect- size was significantly larger than the missense burden effect-size in the *SORL1* and we observed similar trends in *ABCA7* and *ABCA1* (**Table S9**, Supplementary Methods). Of note: missense variants in *ZCWPW1* did not contribute to the most significant burden but were shown here for reference purposes (REVEL>25).