PLOS GENETICS

RESEARCH ARTICLE

Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences

Lingyu Zhan[1]*, **Jiajin Li**[2], **Brandon Jew**[3], **Jae Hoon Sul**[4]*

1 Molecular Biology Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America, **2** Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America, **3** Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, California, United States of America, **4** Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California, United States of America

* zhanly812@g.ucla.edu (LZ); jaehoonsul@mednet.ucla.edu (JHS)

Citation: Zhan L, Li J, Jew B, Sul JH (2021) Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences. PLoS Genet 17(9): e1009772. https://doi.org/10.1371/journal.pgen.1009772

Editor: Amanda J. Myers, University of Miami, Miller School of Medicine, UNITED STATES

Received: April 11, 2021

Accepted: August 10, 2021

Published: September 13, 2021

Data Availability Statement: "The whole-genome sequencing (WGS) and expression data (RNA-Seq) underlying the results presented in this study are available from third parties. Specifically, the ADSP case-control and family WGS datasets are available from the NIAGADS database (accession number NG00067; citation: Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, et al. The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet. 2017;3 (5):e194.; link: https://dss.niagads.org/) through application. The AMP-AD WGS and expression

Abstract

Late-onset Alzheimer's disease (LOAD) is the most common type of dementia causing irreversible brain damage to the elderly and presents a major public health challenge. Clinical research and genome-wide association studies have suggested a potential contribution of the endocytic pathway to AD, with an emphasis on common loci. However, the contribution of rare variants in this pathway to AD has not been thoroughly investigated. In this study, we focused on the effect of rare variants on AD by first applying a rare-variant gene-set burden analysis using genes in the endocytic pathway on over 3,000 individuals with European ancestry from three large whole-genome sequencing (WGS) studies. We identified significant associations of rare-variant burden within the endocytic pathway with AD, which were successfully replicated in independent datasets. We further demonstrated that this endocytic rare-variant enrichment is associated with neurofibrillary tangles (NFTs) and age-related phenotypes, increasing the risk of obtaining severer brain damage, earlier age-at-onset, and earlier age-of-death. Next, by aggregating rare variants within each gene, we sought to identify single endocytic genes associated with AD and NFTs. Careful examination using NFTs revealed one significantly associated gene, *ANKRD13D*. To identify functional associations, we integrated bulk RNA-Seq data from over 600 brain tissues and found two endocytic expression genes (eGenes), *HLA-A* and *SLC26A7*, that displayed significant influences on their gene expressions. Differential expressions between AD patients and controls of these three identified genes were further examined by incorporating scRNA-Seq data from 48 post-mortem brain samples and demonstrated distinct expression patterns across cell types. Taken together, our results demonstrated strong rare-variant effect in the endocytic pathway on AD risk and progression and functional effect of gene expression alteration in both bulk and single-cell resolution, which may bring more insight and serve as valuable resources for future AD genetic studies, clinical research, and therapeutic targeting.

data can be accessed from the AD Knowledge Portal (Synapse: syn3388564; syn10901595; syn3157322; citation: De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142.; citation: Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016;3:160089.; citaiton: Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5:180185.; Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci. 2019;22(12):2087-97.; link: https://adknowledgeportal.synapse.org/Explore/Programs/DetailsPage?Program=AMP-AD) through applications. Besides these underlying WGS and expression data, all other data and information generated in our study are fully available without restriction in the manuscript and supporting information files."

Funding: This work (by the corresponding author, JHS) was supported by the National Institute of Environmental Health Sciences (NIEHS) [K01 ES028064] (https://www.niehs.nih.gov/careers/research/trainingfrom/career/k01/index.cfm); the National Science Foundation grant [#1705197] (https://www.nsf.gov/funding/); the National Institute of Neurological Disorders and Stroke (NINDS) [R01 NS102371] (https://www.ninds.nih.gov/Funding/About-Funding/Grant-Mechanisms); and NINDS [R03 HL150604] (https://www.ninds.nih.gov/Funding/About-Funding/Grant-Mechanisms). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Late-onset Alzheimer's disease (LOAD) is the most common type of dementia and a leading cause of death in the world. Clinical and genetic studies have suggested the potential contribution of the cellular transportation pathway to AD with an emphasis on common variants. In this study, we investigated the effect of rare variants within the cellular transportation pathway and examined three large datasets with over 3,000 individuals with European ancestry. We reported enrichment of rare deleterious variants in the cellular transportation pathway in AD patients from all three datasets. We also observed an elevation of rare deleterious variants in this pathway was associated with individuals with severer brain damages (AD progression), earlier age-at-onset, and earlier age-of-death. By aggregating rare variants in each gene from the cellular transportation pathway, we revealed one gene in which rare variants were significantly associated with the progression of AD. By integrating gene expression data from brain tissues, we identified two additional genes whose rare-variant effect displayed significant influences on gene expression. Taken together, our results demonstrated that rare-variant effect in the cellular transportation pathway is strongly associated with the risk and the progression of AD, which may serve as future clinical and therapeutic targets.

Introduction

Alzheimer's disease (AD) is a destructive and irreversible neurodegenerative disorder, predominantly targeting the elderly.[1] It accounts for 60–70% of dementia cases, characteristic of progressive disintegration of cognitive functions, language ability, and memory loss.[1,2] Late-onset Alzheimer's Disease (LOAD) is a subcategory of AD that appears in persons aged 65 years or older, showing a greater incidence rate as age increases.[3] As the population of Americans age 65 and beyond is expected to reach 88 million by 2050, the number of new AD cases is predicted to double and the prevalence rate to quadruple [4,5].

AD is known to have a substantial genetic component with multiple modulating genes. One of the strongest risk factors for LOAD is *APOE*. Recent GWASs have identified over 50 risk loci accounting for, together with all common SNPs, over 33% of the overall estimated heritability [6–12] that cohered into three major AD-related biological pathways: the cholesterol metabolism pathway, the immune response pathway, and the endocytic pathway.[13] While AD studies have mostly focused on the effect of common variants, such as in the lipid metabolism and immune system/response pathways implicated in recent GWASes, rare variants in genes related to these pathways have not yet been thoroughly investigated.[11,12,14–20] Among these implicated pathways, the endocytic pathway has been identified as one of the most prominent targets, where the earliest morphological changes can be observed as endosome enlargement in post-mortem brains from sporadic AD patients, as well as in some familial cases.[21,22] This phenomenon can be viewed as nearly diagnostic precision and served as blood-cellular markers.[23,24] These findings have also been supported by a recent genetic study showing the enrichment in clathrin-mediated/early endocytosis [25] and clinical research on the facilitation of Aβ clearance by LC3-associated endocytosis.[26] Previous studies using common variants have also identified several risk loci in the endocytic pathway, including *BIN1*, *PICALM*, *CD2AP*, *EPHA1*, and *SORL1* [27].

However, despite being one of the histological hallmarks of AD, few studies have examined the effect of rare variants within this endocytic pathway on AD pathogenic progression.[13] It

is thus of interest to study the rare-variant effect on AD in this pathway. One major challenge in the rare variant study is the lack of power due to their rarity. In this study, to overcome this issue, we analyzed large-scale whole-genome sequencing (WGS) datasets that were recently developed for the study of AD-related traits, including the Alzheimer's Disease Sequencing Project (ADSP) and the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD). Another efficient tool we leveraged to increase the power was a gene-set burden analysis, where we focused on the collective rare variant effect within a set of genes of a known biological pathway, rather than the effect of single variants or single genes, and thus avoided the multiple testing burden required otherwise. This method has helped identify risk genes in various complex traits, such as in central nervous system pathways of schizophrenia.[28–39] In some studies, this method has led to the discovery of novel biological pathways and therapeutic targets through the identification of gene networks participating in the same functional processes.[40–45] Similar gene-set analyses focusing on biological pathways, as well as gene-ontology-based pathway/module analyses, have also been effectively demonstrated in AD studies [11,46,47].

Therefore, in the current study, we included three large-scale WGS datasets with a total of 3,255 individuals of European ancestry, meta-analyzed under a gene-set rare-variant burden analysis framework. Phase 1 of this framework aimed to explore the effect of rare variants in the endocytic pathway as a whole and consisted of two stages followed by meta-analysis. Besides AD status, we additionally explored three AD-related phenotypes, neurofibrillary tangles (NFTs), age-at-onset (AAO), and age-of-death (AOD), along with the phase 1 analysis. NFT status was measured as Braak stages, first proposed by Braak and Braak in 1991, and served as a histopathological indicator of AD, [48–50] representing a finer progression of AD. Phase 2 of this framework was to identify single endocytic genes driving the rare-variant association we captured in phase 1. For each dataset, we examined each gene in the endocytic pathway using both AD and NFT status, followed by meta-analysis across all datasets. Finally, in phase 3, we sought to explore the functional consequences of the rare-variant effect identified in previous phases by examining both the bulk and single-cell expression of endocytic genes in relationship with AD status.

Methods

Study sample

To identify AD-associated rare-variant effects, we evaluated three publicly available large-scale WGS datasets collected for LOAD patients, downloaded as multi-sample VCF files. The Alzheimer's disease sequencing project (ADSP) Umbrella is a collection of sequencing data from the ADSP and other AD and Related Dementia studies. Under this Umbrella, the ADSP group sequenced a large number of well-characterized Alzheimer's Disease (AD) patients at three National Human Genome Research Institute Genome Centers (NHGRI) (Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, and the McDonnell Genome Institute at Washington University). The ascertainment methods and inclusion criteria are described in detail on the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS).[51,52] The sequencing results were mapped to the human reference genome (GRCh38) and processed using the VCPA 1.0 pipeline, which follows GATK best-practices pipeline.[53] Details of the variant calling pipeline can also be found on the NIAGADS. The ADSP discovery extension phase sequenced whole genomes of 1,466 cases and 1,534 controls from five cohorts provided by the Alzheimer's Disease Genetics Consortium (ADGC) and included samples with diverse ancestry backgrounds (Non-Hispanic White, Caribbean Hispanic, and African American). Another WGS project shared under the

ADSP Umbrella is the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is a longitudinal multi-center (63 sites across North America) study designed for early detection and tracking of AD. The ADNI WGS data contains 808 participants with 238 AD cases, 322 mild cognitive control (MCI) subjects, and 248 controls. A full list of the ascertainment methods and inclusion criteria can be found in detailed descriptions in the online ADNI protocol.[54] As of 2018, the ADNI was recalled under the same VCPA 1.0 pipeline as the ADSP discovery extension WGS data and mapped to the same human reference genome (GRCh38), which were then released together. This combined ADSP case-control dataset contained WGS data from a total of 3,896 individuals (accessed by us on Nov 20, 2018), which then underwent a sequence of quality control steps discussed later before including in our stage 1 analysis. Detailed demographic information of this dataset can be found in Table 1 and the distribution of age among AD cases and controls in S12 Fig. To note, we removed samples in the MCI category to ensure a strict bipartite definition of disease status from all our analyses.

Our stage 2 replication included 1,894 WGS samples from the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) Target Discovery and Preclinical Validation Project (accessed by us on Dec 13, 2018). The samples were separately sequenced at three centers: the Religious Orders Study and Memory and Aging Project (ROSMAP) (1,200 samples), the Mount Sinai Brain Bank (MSBB) study (354 samples), and the Mayo Clinic Brain Bank (Mayo) (350 samples). Previously reports have the detailed data collection scheme and sample inclusion and exclusion criteria.[55–58] This sequence data were mapped to the human reference genome (GRCh37) and were processed using the GATK best-practices workflow v3.4.0.[58] Another stage 2 replication was performed on the ADSP discovery extension phase family samples, which were released together with the ADSP case-control data. Therefore, this family WGS data were also mapped to the human reference genome (GRCh38) and processed using the VCPA 1.0 pipeline. The ADSP WGS family dataset contains 888 samples from 161 multiplex families. The inclusion criteria prioritized families loaded with LOAD with minimal *APOE* ε4 alleles. A detailed description of the study design and sample ascertainment methods can be found in previous reports.[59,60] For AMP-AD case-control study, this resulted in 642 AD patients and 969 controls after removing low-quality samples. For the ADSP family study, we obtained 545 AD patients and 285 cognitively normal older individuals (Table 1).

Table 1. Summary of clinical, demographic, and technical information of individuals from three large WGS datasets.

WGS Datasets	Case-control		Family
Studies	ADSP	AMP-AD	ADSP
Total sample size	1,291	1,611	353
EUR Population (%)	41%	93%	50%
AD Patients	664	642	209
Controls	627	969	144
Males (%)	53.4%	35.4%	65.7%
APOE ε4 carriers (%)	43.5%	38.5%	44.8%
Reference genome	GRCh38	GRCh37	GRCh38

The numbers were counted only among the samples included in this current study. The percentages of EUR population were based on the total number of samples within each dataset and only the samples with EUR ancestry were included in this study, which served as the total input sample size in the first row. Abbreviations: AD: Alzheimer's disease; WGS: whole-genome sequencing; EUR: European; ADSP: the Alzheimer's disease sequencing project; AMP-AD: the Accelerating Medicines Partnership-Alzheimer's Disease.

https://doi.org/10.1371/journal.pgen.1009772.t001

RNA-Seq data used for functional analysis were also obtained from the ROSMAP study of the AMP-AD Consortium [58]. The bulk RNA-Seq data were generated for 636 samples (254 AD cases, 368 controls, 12 other dementia, and two without annotation) from the dorsolateral prefrontal cortex (DLPFC) tissues by the Broad Institue's Genomics Platform and processed in an automatic and parallelized pipeline.[55] The ROSMAP group also selected 48 post-mortem samples (24 with severe AD pathology and 24 with low-to-no pathology) and conducted droplet-based single-nucleus RNA sequencing of the prefrontal cortex region.[61] Metadata of the RNA-Seq data were then used to map samples to cases and controls following the same rule as in stage 2 replication, as well as to merge with genotyping data.

Data processing and quality control of WGS data

Individual-level quality control of WGS data. We conducted stringent quality control (QC) to ensure that we only include high-quality samples. As the X chromosome was not available in the ADSP datasets, we did not include X chromosome for all analyses. Before checking the sequencing quality of each individual, we first removed variants failing Variant Quality Score Recalibration (VQSR) in the GATK pipeline and set all variants with genotyping quality (GQ) below 21 to missing. We included only bi-allelic variants for all future analyses. Within the remaining variants, for each individual, we evaluated the genotype missing rate, calculated theoretical relatedness to check for unexpected relationships by study design, and performed principal component analysis (PCA) to identify ancestral composition and population outliers. For the individual-level missing rate QC, we set the cutoff at 5% and removed all individuals beyond this threshold. For the relatedness check, we used PLink 1.9[62] and conducted identity by descent (IBD) analysis, which allowed us to compute a relatedness degree for each sample. For case-control studies, we retained only one in each cluster of samples estimated to be first- or second-degree relatives or duplicates within the corresponding cluster. For the ADSP family study, we compared the empirical kinship relationship record to our computed theoretical relatedness. For PCA, we used 1000 Genomes (1KG) phase 3 as a reference panel.[63] We used EIGENSTRAT [64] for PCA and included only independent common SNVs that were shared between 1KG and our dataset. To note, as PCA assumes unrelated individuals, when performing PCA for the ADSP family cohort, we restricted to only one sample in each family to avoid confounding ancestral relationship by kinship relationship. After having determined the ancestry of the included sample based on PCA, we then assigned that ancestry to the entire family of the included sample. PCA plots (PC1 vs. PC2) of all three datasets could be found in S4 and S6 Figs. As the X-chromosome was available for the AMP-AD study, we also performed sex-check for the AMD-AD and obsevered no sex-mismatched samples. In summary, after stringent sample-level quality control and the careful examination of ancestral backgrounds, we identified 1,291 (664 AD cases and 627 cognitively normal older controls), 1,611 (642 AD cases and 969 controls), and 353 (144 AD cases and 209 controls) high-quality European samples in the ADSP case-control, the AMP-AD case-control, and the ADSP family datasets, respectively, which then served as the primary objects of our study in both stages 1 and 2.

Variant-level QC of WGS data. We conducted stringent variant-level quality control to ensure keeping only high-quality SNVs. We included only variants that served as inputs for the individual-level QC while including only samples passing the individual-level QC. For each variant, we assessed the genotype missing rate, computed minor allele frequency (MAF) using all European samples, and calculated the Hardy-Weinberg Equilibrium (HWE) p-values using only unaffected European samples. For the variant-level missing rate QC, we set the cutoff at 2% and removed all variants beyond this threshold. For HWE, we set the cutoff at 0.001 for rare variants and removed all rare variants falling the HWE check where rare variants are

defined in the following section. The number of variants passing the HWE filter could be found in S14 Table.

Identification and annotation of rare variants

To identify rare variants, we used both external and internal sources of allele frequency to avoid potential inflation of the allele frequency introduced by the study design. For the external sources, we looked at the Europeans (EUR) in 1KG[63] and Non-Finnish Europeans (NFE) in the gnomAD v2 database [65], which matched the ancestral backgrounds of our datasets. We used two different MAF thresholds (0.1% and 1%) to define rare variants, as there is no one consensus definition of rarity and we will correct for testing multiple MAF thresholds in future analysis. In practice, when a variant was present in either of 1KG EUR or gnomAD v2 NFE samples and below the aforementioned threshold, we would keep it for further analysis. For the internal sources, we retained only samples with European ancestry based on the previous PCA, as different ancestral groups would have different allele frequency distributions. Then when a variant was absent in both external databases, we would look at the MAF estimated from the European samples within our dataset and selected rare variants based on 0.1% and 1% MAF thresholds separately. We then annotated rare variants using Ensembl Variant Effect Predictor (VEP)[66]. We defined a variant to be 'deleterious' if it is within one of the following categories: stop-gain, stop-loss, frameshift, splice-donor, splice-acceptor, and missense variants. Particularly, for missense variants, we additionally consulted PolyPhen-2[67] and retained only confident missense variants predicted to be 'damaging.' This definition of deleteriousness focused on coding regions, primarily due to the fact that the effect of non-coding variants was challenging to predict.[68,69] A distribution of variant types and singletons among the selected set of rare deleterious variants could be found in S9 Fig and S11 Table. In an additional validation of the deleteriousness, we further introduced the CADD score [70] as a third deleterious criterion in phase 1 analysis combined with VEP and PolyPhen-2. The distribution of CADD scores among the set of rare deleterious variants could be found in S8 Fig. As suggested by the CADD documentation, variants with scaled CADD > 15 were retained as pathogenic variants and the set of rare deleterious/pathogenic variants passing all three annotation tools were used in this validation test.

Identification of genes in endocytic pathways

We identified genes involved in endocytic pathways using AmiGO 2 [71,72] gene ontology database to select all genes participating in this pathway. We identified three specific GO terms related to the endocytic system in the Homo Sapiens category, which corresponded to three specific compartments in the endocytic system (endosome, lysosome, and trans-Golgi network). The endosome compartment is a membrane-bound vacuole in eukaryotic, participating in the endocytic trafficking from the trans-Golgi network to the plasma membrane and vice versa.[73] The trans-Golgi network serves as an interconnected tubular network and the final cisternal structure involved in packaging and transporting of cargos to the lysosome, endosome, and cell surface.[74] The lysosome, a small membrane-bound lytic vacuole, is one of the end-point in the endocytic transporting pathway, which contains hydrolytic enzymes to break down various biomolecules.[75] The combination of these three compartments formed the essential backbone of the endocytic system, which we named as "endo-system" and used this term throughout the paper. After removing duplicates, we obtained 1,435 genes in total in the endo-system, while the three compartmental gene-sets contained 899 (endosome), 678 (lysosome), and 236 (trans-Golgi network) genes, respectively (S16 Table). We confirmed their biological functions with a functional enrichment analysis using the Database for

Annotation, Visualization, and Integrated Discovery (DAVID)[76], where the top enriched GO terms were indeed lysosome, endosome, and trans-Golgi network. (S13 Fig) A comparison of the endo-system gene-set to the findings in the recent AD GWASes [11,12] has been provided by checking the number of endocytic genes implicated in Jansen et al. and Kunkle et al. (S7 Fig). To note, some genes were related to multiple compartmental gene-sets and thus only one of the duplicated genes was included in the endo-system gene-set (S11 Fig).

Analysis of association between the burden of rare deleterious SNVs and AD status

To identify whether rare variants in the endocytic pathway are associated with AD, we compared the burden of rare deleterious SNVs between AD patients and controls. The burden was defined as the fraction of the alternative minor alleles that each individual carried for all rare deleterious SNVs, using the—score function in PLINK [62]. We additionally performed this procedure on the three compartmental gene-sets and obtained a burden score for each individual within each gene-set. To correct for potential confounding factors, for each gene-set, we first regressed the burden against the total number of rare SNVs and the top ten principal components (PCs). Due to randomness, the distribution of the number of rare SNVs might be naturally variable from sample to sample, in which case the distribution of rare deleterious SNVs would also be greatly affected. Similarly, the PCs helped to correct for potential population stratification within European ancestries. Both aspects could influence the burden score in ways unrelated to AD and thus need to be controlled. Once we had removed the confounding covariates, we performed three logistic regression models as proposed by Zhang et al.[77] using the residuals and AD status for all case-control studies. The three models differed in the covariates they corrected for. The minimal adjustment Model 0 (M0) controlled for the ten PCs and sequencing centers. This model has been previously reported to improve power for detecting variants whose effects are confounded with age and sex.[60] This phenomenon could be introduced by study design where the mean age between cases and controls are substantially disproportionate, as in the case of ADSP studies. Model 1 (M1) was built upon M0 by additionally including age and sex. Model 2 (M2) was further built upon M1 and included the count of *APOE* ε2 and ε4 alleles. For the ADSP family dataset, we generated kinship matrices and used a generalized linear mixed model (GLMM) to take kin relationships into consideration when calculating association p-values. In particular, we used the glmmkin function in the R package, GMMAT.[78] We computed odds ratio (OR) and p-values of association between the burden of rare deleterious SNVs and AD status in each model for European samples in each dataset (ADSP case-control study, AMP-AD case-control study, and ADSP family study). Our stage 1 analysis involved only the ADSP case-control dataset as the discovery set, while the AMP-AD case-control and the ADSP family study served as replication sets in our stage 2 analysis. We chose this analysis scheme because the ADSP case-control study encompassed the largest sample size, including non-European samples, even though we identified fewer samples with European ancestry compared to the AMP-AD case-control study. To note, the AMP-AD case-control study provided only the age-of-death for each individual, while the ADSP case-control and family studies provided only the age-at-onset. As a result, we used different definitions of age in analyzing different datasets. To validate our gene-set AD association analysis, we tested two additional methods provided by MAGMA [79] using the same set of rare deleterious variants. The first was the SNP-wise method applicable to both common and rare variants and the second was the burden method that MAGMA suggested to use for rare-variant-only analysis and was similar to the aforementioned gene-set AD association analysis using PLINK. We applied both methods to the set of rare deleterious variants

previously defined and computed two types of p-values: a competitive p-value that tests whether the association within the gene-set is greater than in other genes and a self-contained p-value that tests whether there is an association within the gene-set of interest at all. The latter concept is the same as what our main analysis method aimed for. Due to our study design with multiple gene-sets and MAF thresholds, a Bonferroni correction was applied in accordance with the number of tests we performed in each analysis to define the study-wide significance threshold in each stage and each dataset. Although our analysis started with the whole endocytic pathway and then moved onto individual compartments, we, nonetheless, utilized a stringent multiple-testing correction threshold. Specifically, as we tested for four gene-sets (endosystem gene-set and three sub-compartmental gene-sets) and two MAF thresholds (1% and 0.1%), we set our significant threshold at $\alpha = 0.05/8 = 0.00625$ for both stage 1 discovery phase and stage 2 replication phase analyses. Accordingly, we set our nominal significance threshold at $\alpha = 0.05$.

To combine results from two stages (three studies) for each of the four gene-sets we tested previously, we performed meta-analyses on p-values using estimates from our best model, namely the model producing the smallest p-values among the three models tested. We used two meta-analysis methods to combine the results. The first was a fixed-effects inverse variance weighted method in METAL [80], which took ORs, standard deviations (SDs), and p-values for separate tests and combined them into one 'Gene-set level' p-value with an estimate of the unified effect. The second was Fisher's method which only required p-values and has been shown to be more robust to some situations where a small portion of p-values are very small. [81,82] In particular, we used the sumlog function from the R package, 'metap,'[83] which took into account the direction of effects in each study and the corresponding p-values. It then computed a 'Gene-set level' p-value similar to METAL indicating the significance of rare variants' effect shared across studies but without an estimated effect size.

Analysis of association between the burden of rare deleterious SNVs and AD-related phenotypes

To test for association between the burden of rare deleterious SNVs and NFTs, we leveraged the Braak stages and followed a similar workflow as in testing AD status. As the sample size of patients with Braak staging information was limited in the ADSP family study, we tested for replication only in AMP-AD case-control study after analyzing the ADSP case-control study in stage 1. We obtained 626 and 1,399 individuals with Braak staging information in ADSP and AMP-AD case-control datasets, respectively. To note, even though the ADSP case-control study had fewer samples with Braak staging information, we, nonetheless, followed the same analysis scheme as in the previous AD analysis. In practice, after removing confounding effects from the burden score, we applied three ordinal logistic regression (OLR) models (M0, M1, M2) to account for multiple ordered categories present in the Braak staging (stage 0 to VI). The regular logistic regression only allows binary dependent variables, which is not feasible for Braak stages. In particular, we used the polr function from the R package, MASS [84], which fits a logistic regression model to an ordered factor response. Similar to the previous burden analysis, our M0 accounted for sequencing centers and the top 10 PCs; our M1 additionally controlled for sex and age; finally, our M2 further included the count of *APOE* ε2 and ε4 alleles. For analyses in all datasets, our significance threshold after the multiple-testing correction was still at $\alpha = 0.00625$ because we tested for two MAF thresholds and four gene-sets. Finally, the nominal significance threshold was also at $\alpha = 0.05$. To increase statistical strength and precision in estimating effects [85], we again performed meta-analyses and combined these two independent tests similar to what we did for AD association analyses.

We additionally tested the age-specific risk of rare deleterious SNVs in the endocytic pathway. As aforementioned, the AAO and AOD information was provided by the ADSP studies and the AMP-AD study, respectively, which allowed us to test for two different age-specific risks within each gene-set. Different from AD risk, age-specific risk leveraged the information of age and estimated the association between the age-to-event (survival time) of patients and the rare-variant burden score. Therefore, we adopted a genetic epidemiological framework proposed by Desiken et al.[86], in which a Cox Proportional Hazard Regression (CPHR) was performed to account for age-to-event information. Specifically, we first used the Surv function from the R package, "survival"[87], and computed a survival time for each sample in each dataset. Then, we conducted CPHR using the coxph function from the R package, 'survminer'[88], to estimate the hazard ratio, or the ratio of risk-to-event (onset or death), depending on the input age we used. We performed three CPHR models (M0, M1, and M2) similar to the previous burden analysis on AD status and Braak staging, except that age was not a covariate in either of the three models. Therefore, since we tested for two different MAF thresholds and four gene-sets (though in a stepwise fashion), we set a stringent significant threshold at $\alpha = 0.05/8 = 0.00625$ and our nominally significant threshold at $\alpha = 0.05$ for analyses in all three datasets. Finally, we combined the results of AAO in the same way as we did for AD and NFT association tests. The resulting p-value then indicated the shared rare-variant effect on AAO-specific risk across the ADSP case-control and family studies.

Single-gene analysis

To identify specific genes within the endocytic pathway associated with AD, we extracted rare deleterious SNVs as defined previously for each gene in the endo-system gene-set that were present in European samples for the ADSP case-control, the AMP-AD case-control, and the ADSP family study. Association test was performed for AD status by first building a null model using the SKAT_Null_Model function in the R package, SKAT, [89] followed by running the SKATBinary function using the SKAT-O feature to obtain association p-values for binary traits. We used a full model that included age, sex, sequencing center, the number of *APOE* ε2 and ε4 alleles, and top 10 PCs. To note, we also applied SKAT_Null_Model to the ADSP family dataset without incorporating kinship structure. This procedure could only be valid in the case where the family structure was relatively simple and did not contribute to a large effect in our analysis. By re-running the previous AD burden analysis with and without kinship information, we indeed observed only small deviations between these two tests. Specifically, for the full model of the endo-system gene-set, we observed an OR of 1.34 with kinship structure provided ($p = 0.035$) while we observed a similar OR of 1.36 assuming an independent setup ($p = 0.02$), which indicated that the family structure within the ADSP family study did not influence our analyses to a large extent.

To test for association with Braak stages, we first extracted only European samples with Braak staging information available for each dataset, before extracting rare deleterious SNVs for each gene within the endo-system gene-set. We leveraged the fact that it is a semi-quantitative trait and performed the association test with the SKAT function for continuous traits with the 'optimal' option after building null models as described for testing AD status. In the attempt to remove confounding factors and unbalanced sample distribution for Braak staging association test, we additionally included AD status in null models. Finally, we meta-analyzed variants across datasets and computed 'Gene-level' p-values for AD status as well as Braak staging. We combined genotyping matrices across three datasets for each gene using the R package, MetaSKAT.[90] Specifically, we first transformed our genotyping matrices into an SSD format for a single population and then analyzed all three populations at once using the

function MetaSKAT_MSSD_ALL. This procedure increased the power to analyze the effects of rare variants that are shared across different studies. To correct for testing multiple genes within the endo-system gene-set, we obtained the number of genes we tested in each separate dataset and computed their corresponding Bonferroni corrected significance thresholds. Specifically, for the AD single-gene analysis, we tested 1,195, 1,228, and 683 genes in ADSP case-control, AMP-AD case-control, and ADSP family datasets, respectively, which corresponded to Bonferroni corrected significance thresholds of $\alpha = 4.18^{*}10^{-5}$; $4.07^{*}10^{-5}$; $7.32^{*}10^{-5}$, respectively. In meta-analyses, we identified 642 genes in common and computed a Bonferroni corrected significance threshold of $\alpha = 7.79^{*}10^{-5}$. For the Braak staging single-gene analysis, we retained only rare deleterious SNVs present in samples with Braak staging information available and tested for 1,035 and 1,176 genes for the ADSP and AMP-AD case-control studies, respectively. The corresponding Bonferroni corrected significance thresholds were then computed as $\alpha = 4.83^{*}10^{-5}$ for the ADSP case-control dataset and $4.25^{*}10^{-5}$ for the AMP-AD case-control dataset. When performing meta-analyses, we examined 967 genes in common between these two datasets, which led to a Bonferroni corrected significance threshold of $\alpha = 5.17^{*}10^{-5}$.

Functional analysis on AD

One approach to understanding how the effect of rare variants would influence the risk of AD status is to investigate how they regulate gene expression. A gene with a variation that is associated with its gene expression is called an eGene. Here, we obtained the bulk RNA-Seq data of DLPFC tissues of 636 individuals from the ROSMAP [55] study and performed an association test between the expression of a gene and rare variants in *cis* with the corresponding gene. In particular, for each gene within the endo-system gene-set, we included all variants within gene boundary and additionally all rare variants within 20kb up- and down-stream of the transcription start sites (TSS), which might potentially regulate the expression of a gene through *cis*-regulation, such as the effect of enhancer region. To overcome the problem of low power to detect the effect of single rare variants, we aggregated the effects of all rare variants within as well as near the TSS of each gene. We analyzed this aggregated effect on gene expression using the SKAT function to compute 'Gene level' p-values, while taking into account confounding covariates, including age, sex, sequencing locations, *APOE* ε2 and ε4 alleles, and top 10 PCs. To correct for testing multiple genes, we calculated false discovery rate for all tested genes and used FDR of 0.05 as the q-value threshold, following the suggestions of previous studies. [91,92] Follow-up validation was performed using genes previous identified from the burden and functional analyses, by directly comparing their expression levels between AD cases and controls using student t-test and computing the Pearson correlation between their expression levels and Braak stages. The multiple-testing issue was then addressed using the Bonferroni correction method.

The resolution of bulk RNA-Seq data may limit our capability of observing cell-type specific effects on AD.[55,61,93,94] To elucidate the underlying complexity of variation across cell types, we further obtained single-cell RNA-Seq (scRNA-Seq) of 48 samples (24 AD patients and 24 cognitively normal controls) from the ROSMAP study and investigated the pattern of expression for each of the six major cell types defined on a priori cell-type-specific gene-sets: excitatory neuron (Ex), inhibitory neuron (In), astrocyte (Ast), oligodendrocyte (Oli), oligodendrocyte-precursor-cell (Opc), and microglia (Mic)[61]. The six major cell types were further divided into sub-clustered cells based on the heterogeneity of gene expression within each cell type: 13 Exs, 12 Ins, 4 Asts, 5 Olis, 3 Opcs, and 4 Mics [61]. The whole dataset in 10X format was first processed using the R package, Seurat.[95] We followed the preprocessing steps as proposed by the Seurat developer by first filtering out cells with reads quantified for less

than 200 or more than 2,500 genes, followed by filtering out cells with the percentage of mitochondrial gene counts over 5 percent. We then employed a global-scaling normalization method provided by the LogNormalize function, which normalized the feature expression measurements for each cell by the total expression, followed by a log-transformation. The major and sub-cell types were identified a priori for this scRNA-Seq data. Therefore, we extracted all significant genes identified in the previous single-gene and functional analyses for each specific cell type and conducted differential gene expression analysis using the student t-test method between cases and controls for each major cell type, as well as for each subcellular population within each major cell type.

Result

The burden of rare deleterious SNVs in endo-system gene-set for ADSP case-control study

To investigate whether rare deleterious SNVs in the endocytic pathway were associated with AD, we leveraged a gene-set method of burden analysis that collapsed individual effects of multiple variants into one 'gene-set level' effect, hence increasing the power of detecting rare variants' effect. We defined rare SNVs using both an external source of allele frequency and allele frequency observed in 1,291 European samples (664 AD cases and 627 controls) from the ADSP case-control study (see Methods). We focused on deleterious SNVs as defined in Methods, in which most were protein-altering variants. We identified rare deleterious SNVs in 1,133 of the 1,435 genes in our gene-set (see Methods). For each individual, we computed the burden of these rare deleterious SNVs. We then compared the genetic burden between AD cases and cognitively normal controls, while taking into account confounding covariates that can potentially influence the amount of burden. Such covariates include ancestral principal components, age, sex, the sequencing location, the number of *APOE* e2 and e4 alleles, and the total number of rare SNVs of each individual. The last procedure is necessary to account for individual differences in the total amount of variation; an individual is likely to carry more rare deleterious SNVs if she/he carries more rare SNVs overall. To note, we found that the total number of rare SNVs on the genome-wide scale has no statistically significant difference between cases and controls ($p = 0.67$, student t-test). As described in Methods, we applied three logistic regression models to find associations between AD status and the burden scores while the three models were built on top of each other and tested for two MAF thresholds (1% and 0.1%). Looking at our best model in terms of the strongest association, we observed that the risk of AD, as indicated by the odds ratio (OR), increased by 1.24 for every one unit increase in residual burden score ($p = 0.00018$ using GLM), which was a significant association after stringent multiple testing correction ($\alpha = 0.00625$) for all gene-sets (including sub-gene-sets we analyzed in next steps) (Fig 1).

Additionally, we identified three major cellular compartments participating in the endocytic pathway and their corresponding genes, which constituted subsets of the endo-system gene-set. The first two compartmental gene-sets were endosome (n = 811) and lysosome (n = 620) gene-sets, which served as the major sorting station in the endocytic pathway and the final destination of proteolytic destinations [96], respectively. The third important compartment was the trans-Golgi network gene-set (n = 208) which represented a pathway sorting station for retrograde trafficking. In summary, we identified 689 endosomal genes, 544 lysosomal genes, and 181 trans-Golgi network genes, respectively. We found the burden scores of rare deleterious SNVs were higher in cases than in controls for all three sub-gene-set. In our best model, the OR, representing the risk of AD, increased by 1.18 per unit for the endosome gene-set ($p = 0.0056$ using GLM), 1.08 per unit for the lysosome gene-set ($p = 0.09$ using GLM;

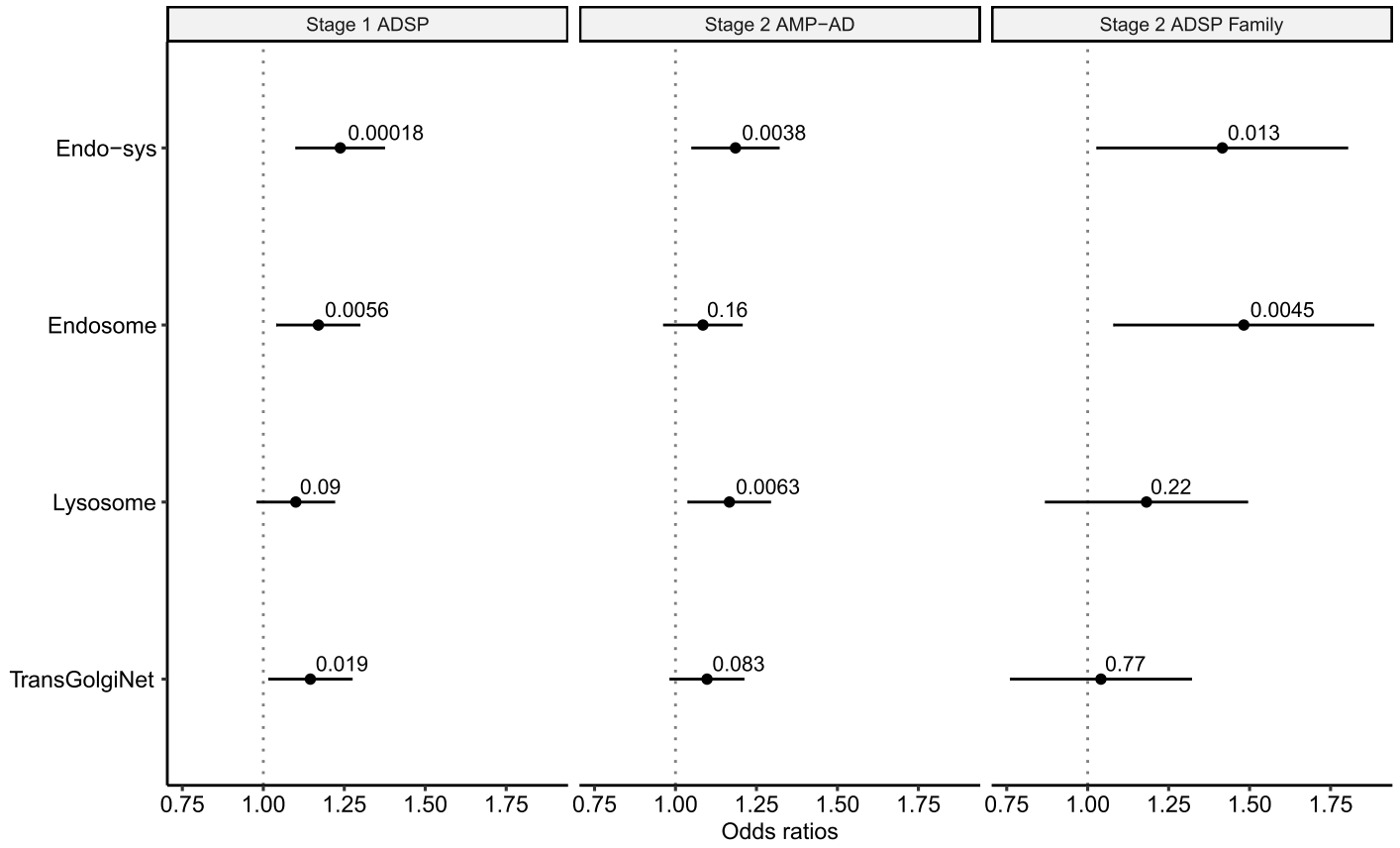

Fig 1. Rare deleterious variants are enriched in AD patients across the endocytic and corresponding compartmental gene-sets in stages 1 and 2. We compared the burden of rare deleterious variants between AD patients and controls across the endo-system (endo-sys) gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (leftmost), which were then tested for replication in stage 2 AMP-AD case-control (middle) and ADSP family (rightmost) datasets. Enrichment (ORs) and p-value were computed using a linear regression model controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1009772.g001

Fig 1), and 1.14 per unit for the trans-Golgi network gene-set (p = 0.019 using GLM). After the multiple-testing correction, we observed the endosome gene-set showed a gene-set-wide significant association with AD while the trans-Golgi network displayed a nominally significant association signal. In addition to exploring sub-gene-sets, we also checked the specificity of the association in the endocytic pathway by obtaining gene-sets unrelated to AD. Specifically, we explored two non-disease complex traits, BMI and height, and obtained related genes (212 and 78, respectively) from GeneRIF, a publically available database for functional annotations.[97] Indeed, we did not observe an enriched rare-variant burden in AD cases compared to controls in these gene-sets and the directions of effects were different across datasets, suggesting the observed rare-variant effect was specific to the endocytic pathway (S13 Table).

Stage 2 replication of the burden analysis in two independent WGS datasets

The gene-set burden analysis in the ADSP case-control study demonstrated statistically significant enrichment of rare deleterious SNVs in cases in the endocytic pathway, indicating an increase of risk conferring AD. We further examined the endo-system gene-set in 1,611 European samples (642 AD cases and 969 controls) from the AMP-AD study. We obtained 1,198

endo-system-related genes and observed an elevated risk of AD in terms of OR of 1.19 ($p = 0.0038$ using GLM; Fig 1), replicating the observation of a significantly higher burden of rare deleterious SNVs in the stage 1 analysis, using a multiple testing threshold of $\alpha = 0.00625$.

We performed additional gene-set burden analysis on the sub-gene-sets of the functional compartments in the AMP-AD study. We identified 735 endosomal genes, 576 lysosomal genes, and 187 trans-Golgi network genes, respectively. We again observed an increase in AD risk among cases for all three sub-gene-sets. A nearly significant signal was observed in the lysosome gene-set with an OR of 1.17 ($p = 0.0063$ using GLM; Fig 1). For the other two gene-sets, we observed an OR of 1.08 ($p = 0.16$ using GLM; endosome gene-set) and 1.10 ($p = 0.083$ using GLM; trans-Golgi network). None of these gene-sets showed gene-set-wide significant association after multiple testing correction at $\alpha = 0.00625$, although the lysosome gene-set nearly reached the gene-set-wide significance threshold.

As described above, the AMP-AD study consisted of three sub-cohorts and the largest one, ROSMAP, contained around 71.5% of the total sample size. To avoid potential batch effect diluting the association signal, we re-performed the analysis on only the ROSMAP data. In fact, we observed slightly more significant results in nearly all gene-sets, where the endo-system and the lysosome gene-sets both reached gene-set-wide significance threshold. (S2 Table) Overall, the associations were similar between the AMP-AD and ROSMAP data, indicating a relatively low level of batch effect among the three sub-cohorts.

Given the observed risk in stage 1 ADSP case-control study and the stage 2 AMP-AD replication, we further examined the genetic burden in the ADSP Family study. We filtered and annotated rare deleterious SNVs based on the same workflow using 353 European samples (144 AD cases and 209 controls) of the ADSP family study. Due to the smaller sample size compared to the previous two case-control studies, we obtained 683 endo-system-related genes. To examine the AD risk, we performed GLMM using the burden of each individual. Due to family structure, we utilized the generalized linear mixed model to account for the relatedness between samples. We observed an OR of 1.42 ($p = 0.013$ using GLMM), conferring an elevated AD risk among cases compared to controls. (Fig 1) This observation was not gene-set-wide significant using the Bonferroni correction threshold at $\alpha = 0.00625$. However, it displayed a nominally significant association with the same direction of effect as in the ADSP and AMP-AD case-control studies.

Nonetheless, we looked into the sub-gene-sets of the three functional compartments in the ADSP family dataset. We identified 402 endosomal genes, 342 lysosomal genes, and 106 trans-Golgi network genes, respectively. We observed a significant elevation of AD risk among cases for endosome gene-set with an OR of 1.48 ($p = 0.0045$ using GLM). Similar increases were also observed in the lysosome and trans-Golgi network gene-sets, with OR of 1.18 ($p = 0.22$ using GLMM) and 1.04 ($p = 0.77$ using GLMM), respectively (Fig 1). Only the endosome gene-set remained gene-set-wide significant after multiple-testing correction, which was in concordance with our observation in the stage 1 ADSP case-control study.

A meta-analysis of stage 1 and 2 burden analysis

The stage 1 burden analysis using the ADSP case-control study demonstrated a significant increase in AD risk in the endo-system gene-set, which was replicated in one independent dataset, the AMP-AD case-control dataset, and displayed a nominal significance in the ADSP family study. We meta-analyzed the results using two different methods (see Methods) and computed a 'Gene-set level' p-value of $2.17^{*}10^{-7}$ (by METAL; Fisher's method produced similar results; Table 2) for the endo-system gene-set, which was improved compared to stages 1 and 2. The same was also observed for sub-gene-sets where we computed a meta-analysis

Table 2. Meta-analysis of stages 1 and 2 gene-set burden analyses using AD, NFT, and AAO.

Phenotype	AD		NFT		AAO	
	P	P*	P	P*	P	P*
Endo-system	2.17E−07	2.66E−07	1.16E−02	9.89E−03	2.47E−06	4.93E−07
Endosome	9.68E−05	6.05E−05	1.30E-01	9.34E-02	3.33E−05	2.04E−05
Lysosome	9.83E−04	1.15E−03	6.56E−03	6.11E−03	1.10E−02	3.11E−04
TransGolgiNet	1.20E−02	7.46E−03	5.71E-01	3.53E-01	2.10E−02	4.96E−03

Abbreviations: AD: Alzheimer's disease; NFT: neurofibrillary tangle; AAO: age-at-onset

NFTs were analyzed using Braak stages. Gene-set-wide significant results were highlighted in bold. Displayed results of gene-set burden analyses were each meta-analyzed using METAL (P) and Fisher's method (P*) (see Methods). Directions of effects were consistent across all tests.

https://doi.org/10.1371/journal.pgen.1009772.t002

p-value of $9.78^{*}10^{-5}$ for the endosome gene-set, $9.83^{*}10^{-4}$ for the lysosome gene-set, and $1.19^{*}10^{-2}$ for the trans-Golgi network gene-set. Except for the trans-Golgi network gene-set that has the smallest number of genes, all other gene-sets remained gene-set-wide significant after multiple-testing correction ($\alpha = 0.00625$), which strongly demonstrated a shared effect of rare deleterious variants within the endocytic pathway across multiple independent studies. To note, although we meta-analyzed the results from the best models, as proposed by Zhang et al. to improve the power of detection, the same pattern of rare-variant association could be observed using the same models for each gene-set across the three datasets. (S1 Table) For all models in the endo-system, endosome, and lysosome gene-sets except M2 of lysosome, we observed gene-set-wide significant p-values, regardless of the meta-analysis methods used, demonstrating a high consistency with the observations made using the best models.

A similar pattern of meta-analysis results was also observed in the additional validation tests from two aspects. Firstly, we wanted to check our results using different annotation tools. Given the set of deleterious variants used in previous phase 1 analyses, we additionally filtered by CADD scores (see Methods) and re-ran the gene-set AD association analyses with the resulting set of pathogenic/deleterious variants. In the meta-analysis, we observed that the endocytic, endosome, and lysosome gene-sets reached gene-set-wide significance threshold (see S5 Table), consistent with the rare-variant effect we observed in the endocytic pathway using the original set of rare deleterious variants.

The second aimed to validate our gene-set burden analysis using MAGMA with two different aggregation methods (see Methods). In the meta-analysis, both the SNP-wise and burden methods provided gene-set-wide significant self-contained p-values for nearly all gene-sets (S3 and S4 Tables; for endo-system, SNP-wise: $9.28^{*}10^{-7}$; burden: $5.16^{*}10^{-8}$), similar to the results shown above in Table 2. Compared to the MAGMA burden method, the SNP-wise method was not designed for rare-variant-only analysis and indeed showed weaker association signals. Especially for the competitive p-values, we observed gene-set-wide significant results for nearly all gene-sets using the MAGMA burden method, but not the SNP-wise method (for endo-system, SNP-wise: $2.41^{*}10^{-2}$; burden: $1.90^{*}10^{-3}$). We also attempted to compute a weighted burden score using pLI scores by PLINK and observed gene-set-wide significant associations in the endo-system gene-set in the meta-analysis. (S10 Fig, S12 Table) Compared to our main method above, the MAGMA methods and the weighted method displayed some fluctuations in individual datasets and models but consistent results in meta-analysis, indicating a robust rare-variant effect in the endocytic pathway under different statistical methods. Besides, as *APOE* was a major risk determinant in AD, in this validation, we also checked whether our observed rare-variant enrichment was mainly contributed from this gene, rather than the whole endo-system gene-set, by re-run the analysis with *APOE* excluded. Indeed, we observed

nearly the same p-values in the meta-analysis, indicating a rare-variant enrichment in AD cases even without *APOE*.

The burden of rare deleterious SNVs on NFTs

NFT, measured in Braak staging, was one of the most important histopathological indicators of AD [48–50]. It is designed as an ordinal scale from 0 to VI of NFT pathology where AD patients with high Braak stages (V or VI) are diagnosed with high confidence.[98] Therefore, Braak stages may serve as a finer spectrum or proxy of AD severity and provide higher power in assessing the effect of rare variants in AD progression. Based on our previous AD analysis, we hypothesized that the burden of rare deleterious variants in the endocytic pathway would be higher in patients with later Braak stages. To test our hypothesis, we applied an ordinal logistic regression (OLR) method to Braak stages (see Methods). This method has been previously shown to be effective in studies of Braak staging as well as of other ordered phenotypes, such as oral cancers.[99,100] We obtained 626 individuals (475 AD cases and 151 cognitively normal controls) from the stage 1 ADSP case-control dataset and 1,399 individuals (533 AD patients and 866 controls) from AMP-AD case-control dataset with Braak staging information, which were used to fitted OLR models. In stage 1, We observed an OR of 1.16 ($p = 0.039$ using OLR; S1 Fig) in the endocytic pathway, implicating a nominally significant association of rare-variant enrichment to later Braak stages. However, this result did not replicate in stage 2 with sufficient significance (OR = 1.08, $p = 0.13$ using OLR; S1 Fig). Comparing the stages 1 and 2 samples, we observed a distinct distribution of Braak stages. In particular, the stage 2 samples were concentrated in Braak stage III (23.1%), IV (28.1%), and V (23.3%), whereas most stage 1 samples were clustered in stage V (26.4%) and VI (34.8). (S3 Fig) We did not test for replication in the ADSP Family study due to limited samples with Braak staging information (n = 38 individuals where only one sample had AD).

Our analyses of two independent datasets suggested a trend of increased risk of bearing later Braak stages with elevated rare-variant burden in the endocytic pathway. To improve power, we meta-analyzed the results from the ADSP and AMP-AD case-control studies, producing a 'Gene-set level' p-value between 0.0099 and 0.012, which did not pass our multiple-testing correction threshold of $\alpha = 0.00625$. (Table 2) Further Braak staging burden analysis using compartmental sub-gene-sets, however, revealed a gene-set-wide significant signal in the meta-analysis for lysosome gene-set ($p = 0.0066$, Fisher's method). A full list of results for NFT burden analysis can be found in S6 Table.

Hazard analysis on population risk of AD age of onset and death

Previous gene-set burden analyses have demonstrated a significant correlation between the burden of rare deleterious variants within the endocytic gene-set and AD risk. One important aspect of AD development is its age-specific phenotypes, such as AAO. Previous studies on AD have shown a large genetic component in the heritability of AAO [101,102], with multiple risk loci associated with it. [103–107] It is thus of interest to also examine the genetic risk identified within the endocytic gene-set in this context. One approach is to evaluate whether AD patients with earlier AAO are associated with greater rare-variant burden within the endocytic gene-set. Previous studies have proposed a genetic epidemiological framework, where age-specific phenotypes were analyzed using a Cox Proportional Hazard Regression (CPHR) that considered a time-to-event probability, as opposed to the simple event probability estimated in logistic regression.[86,108] Therefore, we leveraged our previously computed burden score for each individual in the ADSP case-control study and constructed a cox proportional hazard (CPHR) model to estimate the instantaneous risk of developing AD, in consideration of

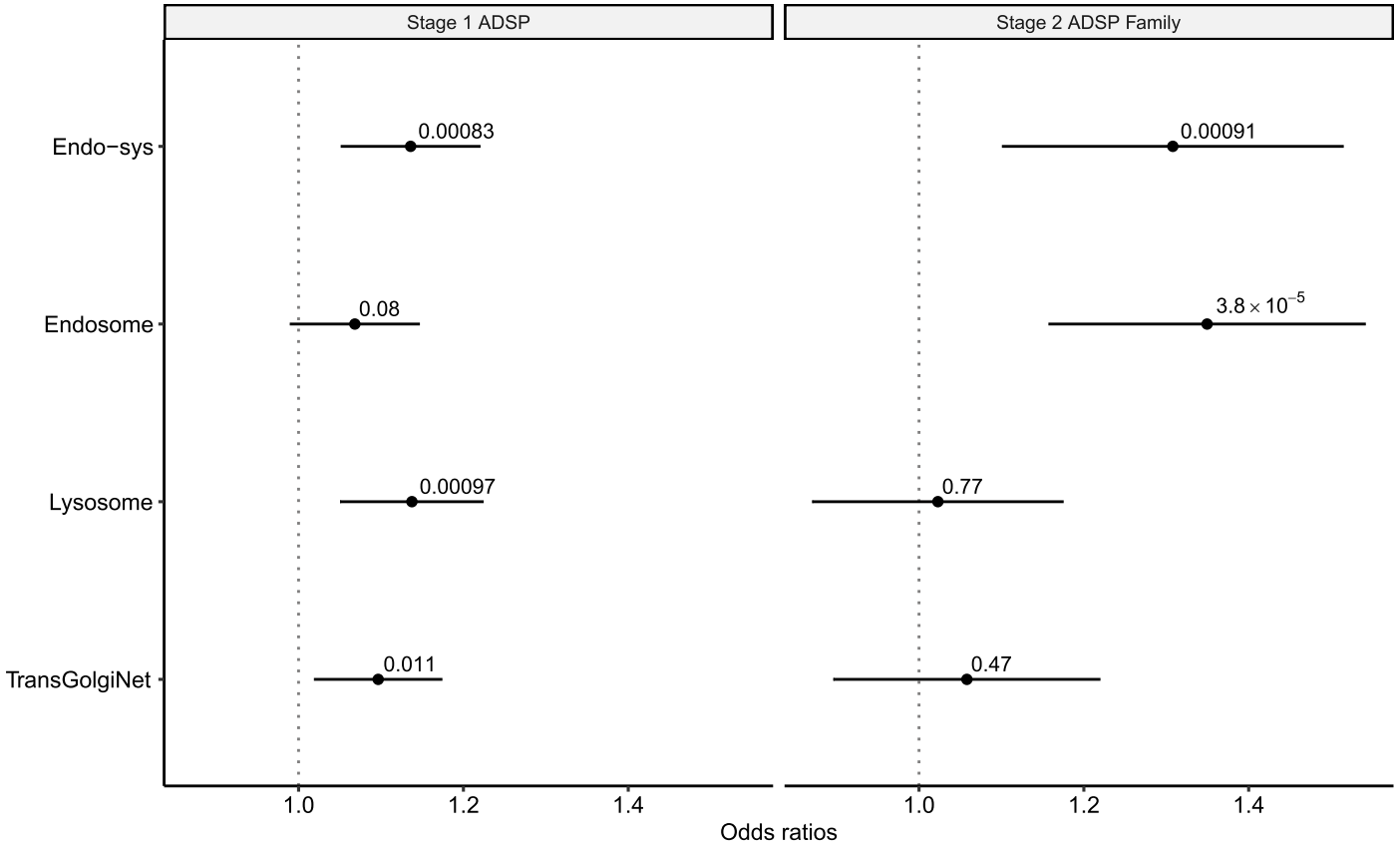

Fig 2. The enrichment of rare deleterious variants is associated with AD AAO across the endocytic and corresponding compartmental gene-sets in stages 1 and 2. We computed a hazard ratio of obtaining AD in earlier ages using the burden of rare deleterious variants across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (left), which were then tested for replication in stage 2 ADSP family datasets (right). Enrichment (ORs) and p-value were computed using CPHR. P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.

https://doi.org/10.1371/journal.pgen.1009772.g002

genotype and AAO. A positive estimate of hazard in this model would indicate a higher risk of developing AD in early ages. We built three models as in the burden analysis and observed in our best model that an AAO-specific genetic risk increased by 1.14 per unit increase in the residual burden score (p = 0.00083 using CPHR; Fig 2), which reached gene-set-wide significance after multiple testing correction ($\alpha = 0.00625$). We further examined the AAO-specific genetic risk within the functional sub-gene-sets. In our best model, we observed a gene-set-wide significant hazard ratio of 1.14 (p = 0.00097 using CPHR) for lysosome gene-set and a nominal significant hazard ratio of 1.10 (p = 0.011 using CPHR) for trans-Golgi network gene-set.

To test for replication, we examined the ADSP family study under the same statistical framework. Applying the CPHR models, we observed a gene-set-wide significant hazard ratio of 1.31 (p = 0.00091 using CPHR; Fig 2) in the endo-system gene-set. Carefully examining the sub-gene-sets also revealed gene-set-wide significant AAO-specific risk within the endosome gene-set (HR = 1.35, $p = 3.83^{*}10^{-5}$ using CPHR). We did not observe significant associations using the other two compartmental gene-sets (S7 Table).

To increase power, we performed meta-analyses to identify rare-variant effects shared across multiple studies. We combined the best results from ADSP case-control

and family studies and observed a gene-set-wide significant p-value of $2.47^{*}10^{-6}$ (by METAL; Fisher's method produced similar results; Table 2) for the endo-system gene-set, which was greatly improved compared to results in either stage. Similarly, the endosome gene-set also demonstrated an improved gene-set-wide significant p-value of $3.33^{*}10^{-5}$. However, the lysosome and the trans-Golgi network gene-sets showed only nominally significant p-values in our meta-analysis, potentially due to the absence of signal in the ADSP family study. These findings strongly demonstrated that this AAO-specific rare-variant effect in the endocytic pathway was shared in European samples across different studies.

Another age-specific phenotype is the age of death (AOD), which has been shown to be affected by genetic groups implicated in AD AAO as well as in other dementia.[109,110] We thus followed the same analysis framework using the CPHR model and assessed whether AD-affected patients with earlier AOD were associated with a higher rare-variant burden in the endocytic pathway. We looked at European samples in the AMP-AD case-control study, where the AOD information was available. We observed a hazard ratio of 1.10 ($p = 0.024$ using CPHR; S2 Fig), indicating an increase of risk of death in AD patients as well as a worse prognosis along with an elevation in genetic burden. Further analysis using the lysosome sub-gene-set displayed a hazard ratio of 1.09 ($p = 0.036$ using CPHR). Both endo-system and lysosome gene-sets demonstrated nominally significant associations with AOD but did not reach gene-set-wide significance after multiple-testing correction. Analysis using other sub-gene-sets did not provide significant hazard ratios.

Single-gene analysis on AD risk using AD and NFT status

From the previous analysis, the endo-system gene-set conferred a large rare-variant effect on AD and related phenotypes. Thus, we decided to examine the effect of rare variants in single endocytic genes, attempting to identify those associated with AD with large effect sizes. To increase power, we aggregated previously defined rare deleterious SNVs in each gene and tested for association with AD. We did not observe a single gene passing the Bonferroni corrected significance threshold in all three datasets, as well as in meta-analysis (See Methods; $\alpha = 4.18^{*}10^{-5}$; $4.07^{*}10^{-5}$; $7.32^{*}10^{-5}$; $7.79^{*}19^{-5}$, for ADSP case-control, AMP-AD case-control, ADSP family studies, and meta-analysis respectively; S15 Table).

As mentioned previously, NFT status may provide more detailed information of the pathological progression of AD and thus a greater power to detect signals of rare-variant effect. We, therefore, performed single-gene analysis using NFT status, as a proxy for AD status. For all datasets, we retained only rare deleterious SNVs that were present in samples with Braak staging information. We controlled for the same set of covariates as in previous analyses, except that we also included the AD phenotype (AD affected/unaffected) for each individual as one additional covariate (see Methods). The latter is necessary because the Braak staging and the AD phenotype are correlated, and the numbers of individuals with and without AD were vastly disproportionate among the samples with Braak staging information. For the ADSP case-control study, we observed six genes that reached Bonferroni corrected significance threshold ($\alpha = 4.83^{*}10^{-5}$). None of the genes passed the Bonferroni corrected significance threshold ($\alpha = 4.25^{*}10^{-5}$) in the AMP-AD study. Results of the top ten most significant genes can be found in S8 Table. We conducted meta-analyses for these two independent studies using MetaSKAT as before. In the combined results, we observed one gene, *ANKRD13D*, reached Bonferroni corrected significance threshold ($p = 3.56e\text{-}05$; $\alpha = 5.17^{*}10^{-5}$). This gene has been previously implicated in AD through RNA expression analysis [111] and protein interactome mapping [112].

The identification of functional effects of rare variants within the endocytic pathway

The hypothesis that the endo-system gene-set contains rare variants that are influential to AD development is endorsed by the previous gene-set burden analyses and single-gene analyses. One approach to investigating how the effect of rare variants takes place is to analyze how these rare variants are associated with gene expression. Such gene containing variations affecting its expression is often called an eGene.[91] To identify eGenes, we obtained bulk RNA-Seq data of DLPFC brain tissues of 636 individuals from the ROSMAP study as part of the AMP-AD study and tested for association of all variants in *cis* with a gene with its gene expression. Specifically, we grouped all variants within one gene, as well as those near the corresponding TSS, and assessed whether the aggregated rare-variant effect in an endocytic gene is associated with its expression level using SKAT (see Methods). Intersecting the bulk RNA-Seq and WGS data revealed 547 individuals with 224 AD patients and 323 controls. By taking an FDR of 5%, we discovered two genes, *HLA-A* and *SLC26A7*, whose rare variants were significantly associated with expressional changes. To note, previous studies have demonstrated that proteins from the same families of these two genes are associated with AD status. Specifically, two proteins from the HLA families and one from the SLC families have been implicated in AD through meta-analyses of large GWAS and brain DNA-methylation association analysis.[9,113] We first examined their single-gene analysis results and observed that none of them was significant using the AMP-AD dataset ($p = 4.74e\text{-}01$; 2.14e-01, for *HLA-A* and *SLC26A7*, respectively). To validate our results and determine the direction of effects, we compared the expression of these two genes between cases and controls. Indeed, their expression levels were both significantly decreased in cases compared to controls ($p = 0.00073$ *HLA-A*; Fig 3A; $p = 0.0054$ *SLC26A7*; Fig 3B; student t-test; $\alpha = 0.017$;). We further examined the distribution of their expression levels across multiple Braak stages. Similarly, both expressions were strongly negatively correlated with greater Braak stages ($r = -0.129$, $p = 0.0024$ *HLA*-A; $r = -0.127$, $p = 0.0029$ *SLC26A7*; Pearson correlation; $\alpha = 0.017$).

We also investigated *ANKRD13D*, which we previously identified to be associated with Braak stages, in the context of gene expression. Although not an eGene, *ANKRD13D* exhibited a significant expressional decrease in cases compared to controls ($p = 0.0026$; student t-test; Fig 3C). The analysis on Braak staging also revealed a strong negative correlation ($r = -0.122$, $p = 0.0042$; Pearson correlation).

scRNA expression analysis

Recent advancement in analyzing gene expression in single-cell resolution has provided opportunities to uncover complex alterations across cell types and identify cell-type specific effects on AD.[55,61,93,94] For example, previous studies have pointed the imbalance of excitatory and inhibitory neurons could lead to overexcitability and early dysregulation in the development of AD [114]. Many other studies also demonstrated abnormalities in innate immune cells, primarily microglia, in the pathogenesis of AD.[115] Therefore, to investigate the potential cell-type specific effects of rare variants within the endocytic pathway, we obtained the single-cell RNA-seq data of 48 samples (24 AD patients and 24 cognitively normal controls) from the ROSMAP study. We focused on three genes we identified through the previous analysis, which demonstrated significant associations to AD progression. The scRNA--Seq data were labeled with six major cell types using a priori marker genes (Ex, In, Ast, Oli, Mic, and Opc), and sub-clustering within each cell type revealed cellular subpopulation (see Methods). We examined the expression of the three target genes in all major cell types and observed that *ANKRD13D* was up-regulated in Ex ($p = 1.92^{*}10–18$; student t-test), Ast

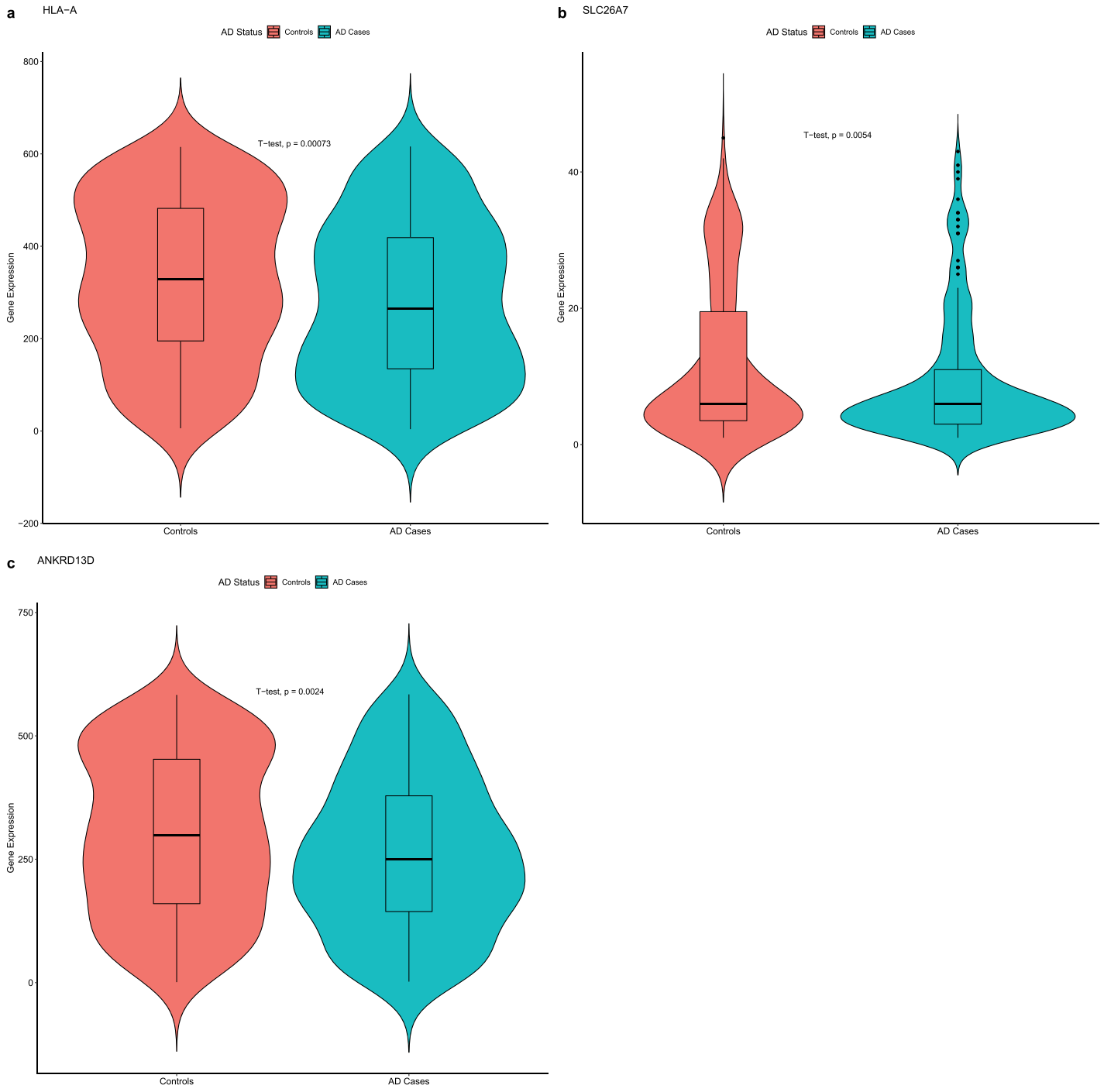

Fig 3. Comparison of the gene expression of *HLA-A*, *SLC26A7*, and *ANKRD13D* between AD cases and controls from the ROSMAP study. Violin plots were used to represent the distribution of gene expression within each AD status, where a symmetric deviation from the middle line on both sides indicated a higher abundance of samples at the corresponding gene expression level. Comparisons between AD cases and controls were assessed using boxplots. P-values were computed using the student t-test. All three genes, *HLA-A*, *SLC26A7*, and *ANKRD13D*, are down-regulated in AD cases compared to controls.

https://doi.org/10.1371/journal.pgen.1009772.g003

(p = 0.011; student t-test), and In (p = 0.028; student t-test) (S9 Table). However, it exhibited a down-regulation in Oli (p = 0.0018; student t-test). *SLC26A7* was observed to be up-regulated in Ex (p = 0.0049; student t-test), while *HLA-A* displayed a pattern of down-regulation in both In and Mic (p = 9.72*10-6and p = 0.0031 respectively; student t-test). Four AD pathology-associated cellular subpopulations (Ex4, In0, Ast1, and Oli0) have been previous demonstrated for this scRNA-Seq data [61,91]. Our differential expression analysis within these four subpopulations showed a pattern of up-regulation of *ANKRD13D* in Ex4 and In0 (p = 5.76*10–8 and p = 0.036, respectively; student t-test; S10 Table). The other two genes, however, were not significantly differentially expressed in these four cell subpopulations.

Discussion

Using large publicly available WGS datasets, our study described here enabled us to assess the contribution of rare variants to AD. In our stage 1 discovery phase, we observed a significantly elevated burden of rare deleterious SNVs in affected individuals compared to cognitively normal older controls within the endocytic pathway. We chose this pathway because it represented one of the earliest morphological changes in AD development, and multiple AD risk factors, predominantly through common SNPs, have been implicated specifically in this pathway with genome-wide significance, including *BIN1*, *CD2AP*, *PICALM*, *RIN3*, and *SORL1*. [9,12,18,19,21,22,116] Our results demonstrated additional correlation between rare variants in the endocytic pathway and AD. Successful replication in the AMP-AD case-control study and improved meta-analysis association further strengthened this contribution of rare deleterious variants to AD risk. Our analysis using the ADSP family dataset showed a similar enrichment of rare deleterious SNVs in AD patients, although not reaching gene-set-wide significance. One possible explanation was that the sample size of this family study was relatively small (one third to one fifth) compared to the other two case-control studies. We additionally identified gene-set-wide-significant signals within the endosome and lysosome gene-sets using meta-analysis, implicating potential compartment-specific roles in AD pathology. One possibility that we did not observe significant results in separate stages for all three sub-gene-sets was because they contained a smaller number of genes compared to the endo-system gene-set and, therefore, smaller aggregated effects of rare variants, which required meta-analysis to combine signals in individual samples. As the smallest gene-set (one-third to one-fourth of the other two), the trans-Golgi network remained nominally significant even after meta-analysis.

In assessing the AD pathological progression, we examined the association of rare-variant effect to NFT pathology using Braak staging. We observed the gene-set-wide significant association within the lysosome gene-set, where individuals with higher Braak stages were enriched with rare deleterious SNVs. No significant association was found in other gene-sets, besides a nominally significant association in the whole endocytic pathway. Compared to the previous analysis using AD status, our analysis using Braak stages was largely limited in sample size. For example, only 626 out of 1,291 European samples in the ADSP case-control dataset had Braak staging information available. For the ADSP family study, only 38 out of 353 samples had Braak information available, which made analyzing Braak stages in this dataset infeasible. Additionally, this was further complicated by the disproportionate distribution of samples across different Braak stages. The ADSP case-control dataset contained 218 samples in stage VI while only 15 samples in stage 0. Such highly skewed distribution reduced our power to detect a significant association between rare variants' effects and Braak stages. The AMP-AD dataset was similarly skewed but also distributed largely differently from the stage 1 dataset. This distinction in distribution may explain why we observed different signals in our stage 1 and 2 analyses.

Based on the idea that rare variants within the endocytic pathway were associated with AD progression, we further tested age-specific phenotypes and leveraged a CHPR model previously proposed to be effective in assessing the effect of variants on age-to-event risk.[86] For AAO, we observed a gene-set-wide significant hazard ratio in the stage 1 analysis, indicating an association of rare-variant burden in the endocytic pathway to earlier AAO of AD, which was replicated in stage 2. A similar observation was found in the compartmental gene-sets, where endosome gene-set demonstrated a gene-set-wide significant signal in meta-analysis. Nonetheless, we did not replicate our stage 1 findings of the lysosome gene-set in stage 2, potentially due to the small sample size of the family dataset and the small size of the gene-set. For AOD, we examined the AMP-AD dataset and only observed nominally significant signals in the endocytic pathway and the lysosomal compartment. Previous analyses on AAO have demonstrated a substantial correlation of AAO between parents and their children, with multiple risk loci, such as *APOE*, *GRN*, *MPT*, and *C9orf72*.[101,109] Genetic studies using AOD from LOAD datasets have revealed additional associations of SNVs in these genes with human aging.[110] Consistent in the observation of significant genetic components, our results discovered an additional contribution of rare variants within the endocytic pathway to age-related phenotypes.

Our discovery of the increased burden of rare-variant effect in AD patients led us to explore the effect of individual genes within the endocytic pathway and attempt to identify specific ones with large effect sizes which might serve as potential clinical and therapeutic targets. We performed single-gene analysis using both AD status and Braak staging as the target phenotypes. When looking at the AD status, we did not observe a gene with a large enough effect to be detected in our analysis. Using Braak staging information, we were able to identify one gene, *ANKRD13D*, that showed robust signal across multiple studies after multiple-testing correction. This may be due to the fact that Braak stages provided a finer indication of AD progression. *ANKRD13D* encodes a member of the Ankyrin repeat domain 13 family, characterized by three ankyrin repeats at the N-terminal facilitating protein-protein interaction.[117] It has been experimentally shown to localize to endosomes and is known to regulate the rapid ubiquitin-dependent internalization and sorting of membrane-bound proteins within the endocytic pathway.[118] One of its main targets is the endocytosis of the epidermal growth factor receptor (EGFR) through the functional ubiquitin-interacting motif (UIM) of the ANKRD13 family proteins, which is then degraded in lysosomes.[118,119] EGFR is a transmembrane protein serving as a receptor epidermal growth factor (EGF) protein ligands. Multiple previous studies have reported abnormal plasma levels of EGF in AD patients, [120–122] and two recent studies on EGF have demonstrated its protective effects on AD by preventing amyloid-beta (Aβ)-induced angiogenesis deficit to brain endothelial cells in vitro and in vivo.[123,124] Recent studies have also described that the EGFR internalization after EGF binding was strongly inhibited when ANKRD13 proteins were over-expressed.[118] This mechanism implicates a potential regulatory effect of the *ANKRD13* family on AD pathology through the regulation of internalization of EGFR. Indeed, the link between *ANKRD13D* and AD is further bolstered by a recent RNA profiling where they identified an altered gene expression of *ANKRD13D* between the blood and brain tissue of AD patients.[111] In our analysis, we identified seven rare deleterious SNVs within *ANKRD13D*, where six were predicted to be missense damaging variants and one was predicted to be either missense damaging or splice region variant. These mutations could potentially alter its ubiquitin-binding ability, either through directly changing the sequence or indirectly through changing the 3D protein folding structure, and affect the normal protective function of EGF in AD development. Further functional studies of *ANKRD13D*, and in particular these seven variants, will be needed to specifically define its role in AD pathogenesis and evaluate the therapeutic and clinical importance of the EGFR pathway.

To investigate the functional effects of rare variants, we looked at the expression of genes in the endocytic pathway at both bulk tissue and single-cell resolutions. Leveraging bulk RNA--Seq data, we identified two significant eGenes, *HLA-A* and *SLC26A7*, in the ROSMAP study. Careful examination of these two eGenes in the context of AD status revealed a pattern of down-regulation in AD patients compared to cognitively normal controls. A similar negative correlation was found using Braak stages. *HLA-A* encodes a member of the human leukocyte antigen A (HLA) class I, also called the major histocompatibility complex (MHC) class I. It has been shown to participate in the important "cross-presentation" mechanism of T cell-mediated immune response, specifically efficient in dendritic cells.[125] This mechanism is part of the endocytic pathway that involves the internalization of HLA class I proteins from the cell surface through early endosomes and the loading of antigen peptides in lysosomes.[126] Previous studies have described an important role of HLA class I in maintaining the integrity of aging brains and have demonstrated significant dendritic atrophy with deficient HLA class I. [127] Moreover, recent GWA studies have identified specific alleles in *HLA-A* associated with AD in the Italian and Chinese population [128,129], as well as risk loci in other members of the HLA family.[9] The other identified eGene, *SLC26A7*, encodes a member of the solute carrier (SLC) family that localizes to subapical lysosomal membrane as well as endosomes, primarily serving as an exchanger and transporter of a broad spectrum of substrates in the endocytic pathway.[130,131] Disruption in the expression of SLC26 proteins has been shown to cause severe acid-base balance dysregulation, leading to disruption of anion homeostasis. [132] Multiple SLCs have been associated with AD, such as *SLC2A2*, which was linked to astrocyte activation leading to its elevation in AD patients, [130] and *SLC1A3*, whose expression has been associated with Aβ deposition [133]. Recent GWA studies have also identified risk loci in members of *SLCs*, such as *SLC24A4*.[9] Specific implication of *SLC26A7* has also been shown through gene co-expression network mining where STAT1, a transcription factor of *SLC26A7*, was differentially expressed between AD patients and cognitively normal controls. [47] In our analysis, we identified nine rare deleterious SNVs in *HLA-A* in which six were predicted to be damaging missense mutations, two were predicted to be splice acceptor variants, and one was predicted to be either damaging missense mutation or splice region variant. In *SLC26A7*, we also identified nine rare deleterious SNVs, which are all damaging missense mutations. As transporters, these two genes could potentially be altered in their affinities to ligands due to changes in primary or tertiary structures. Our results here supported these previous findings and provided additional evidence from the aspect of the rare-variant effect on gene expression. Further investigation will be required to elucidate specific variants conferring these effects as well as other participating proteins in the same signal relay mechanisms of *HLA-A* and *SLC26A7*.

In a single-cell resolution, we further explored the cell-type-specific functional effects of the significant genes identified in our previous analyses. Previous single-cell transcriptomic analyses have shown a large number of cellular subpopulations with cell type-specific associations with AD.[61] Our analysis supported this finding in *ANKRD13D*, *HLA-A*, and *SLC26A7*. For example, we observed an up-regulation of *ANKRD13D* in bulk tissue, but it was found to be regulated differently in different cell types: up-regulated in Ex, Ast, and In, while down-regulated in Oli. On the other hand, in single-cell RNA-Seq data, *SLC26A7* and *HLA-A* showed a pattern of down-regulation in AD patients, consistent with our findings using the bulk RNA--Seq data though with various effect sizes in different cell types.

Several strengths and limitations of our study warrant discussion. One of the major strengths is our study design to begin the analysis with pathways implicated in AD a priori. Our usage of the endocytic pathway provided us the power to identify rare-variant effects that would otherwise be missed in traditional association analysis of single variants. This design

was further combined with the large sample sizes of the three independent datasets, which provided additional power. We separated these datasets into a discovery phase and a replication phase and were able to replicate our discovery phase results in two independent datasets of the replication phase, followed by meta-analyses of samples in all three studies. This procedure ensured us to identify and validate associations while retaining large power to identify small signals. Another strength of our study is the analysis of AD-related phenotypes, such as Braak stages, and provided additional power in identifying single genes with large aggregated rare-variant effect sizes. The analysis of AAO and AOD provided further information on the progression of AD, which is especially important in clinal AD prediction and intervention. One more strength in our analysis lies in our exploitation of bulk- and sc-RNA expression data in combination with AD genotyping data. Through this method, we were able to identify eGenes with large rare-variant effect, which would require a much higher sample size and greater power to be identified as eQTLs and suggested potential AD-regulating mechanisms.

One limitation of the study is that while we used WGS datasets, we only focused on analyzing rare SNVs within genic regions. Our analysis relied on knowing the deleteriousness of each variant contributing to the gene-set burden, and variant annotation is most reliably predicted for coding and splice site variants.[90,134] Including variants in intergenic regions or indels may result in the inclusion of variants with benign effects and decrease our power of detecting AD-associated genetic burden. Another limitation of our study is that even though we utilized WGS datasets of large sample size, they were not large enough to detect single genes where rare variants significantly influenced AD. Although our analyses displayed sufficient power to detect rare-variant effects within sets of genes, we nonetheless failed to directly identify direct gene-level associations with AD. To achieve this latter goal, we may need WGS datasets of larger sample sizes. A similar limitation on sample sizes was seen in those with expression data and Braak staging information. Our bulk RNA-Seq data is only available for 547 individuals from the ROSMAP study in which we have genotyping data for 1200 individuals. The scRNA-Seq data is further limited in that we have 48 samples from the ROSMAP study. These limitations in sample size decreased our capability of detecting functional effects of rare variants within the endocytic pathway. One more limitation in this study is that we primarily focused on European samples because we had a limited sample size for non-European ancestries across all three WGS datasets. Nonetheless, it may be of interest to check whether we would observe similar rare-variant effect in the endocytic pathway in non-European samples as we observed in European samples. Another limitation rooted in the potential batch effects among the ADSP datasets used in this study, as also mentioned in Holstege et al.[135], due to the fact that the samples were sequenced and called in different locations. In this study, we have addressed the potential batch effect from three aspects. Firstly, the version of the ADSP datasets used in this study has been quality controlled, where all samples from different centers were re-processed using the same VCPA 1.0 pipeline and corrected for many technical issues present in the previous version, including contaminations, mismatches, and duplicates. [136] Secondly, we conducted additional QC steps at variant-level and sample-level. These included many steps suggested by Holstege et al., such as sex-check, selecting European samples by PCA, removing unexpected related samples using IBD, checking for samples with aberrant Ti/Tv ratio or novel SNV/indel count, and filtering out variants failing VQSR, GQ, HWE, and missing rate thresholds. Thirdly, we included sequencing location as a covariate in all models (M0, M1, and M2) to account for potential batch effects. Therefore, in this study, we recognized and have carefully approached this limitation, as much as we could, to mitigate the potential batch effects.

In summary, our study demonstrated significant rare-variant effect within the endocytic pathway in European samples. Such effect was also associated with Braak stages and age-

related phenotypes, suggesting a potential target for clinical and therapeutic studies. Further investigation within this pathway revealed one gene significantly associated with Braak stages and two eGenes with a pattern of differential expression between AD patients and cognitively normal controls. More functional studies will be necessary to gain a better understanding of their molecular mechanisms of how they participate in the processing and modification of AD-related proteins. In vitro and in vivo experiments on these genes will also provide further insights into the connections of genetic variants to their gene expression and elucidate protein signaling models that affect the pathogenic progression of AD.

Supporting information

S1 Fig. Rare deleterious variants are enriched in patients with severe NFTs across the endocytic and corresponding compartmental gene-sets in stages 1 and 2. We compared the burden of rare deleterious variants between patients with different severity of NFT across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (left), which were then tested for replication in stage 2 AMP-AD case-control dataset (right). Enrichment (ORs) and p-value were computed using OLR controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.
(EPS)

S2 Fig. The enrichment of rare deleterious variants is associated with AD AOD across the endocytic and corresponding compartmental gene-sets. We computed a hazard ratio of earlier AOD with AD using the burden of rare deleterious variants across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in the AMP-AD study. Enrichment (ORs) and p-value were computed using CPHR controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.
(EPS)

S3 Fig. Distribution of Braak stages in individuals from Stage 1 ADSP and Stage 2 AMP-AD datasets.
(EPS)

S4 Fig. PCA plots (PC1 vs. PC2) of the ADSP case-control dataset showing the distribution of ancestry backgrounds.
(EPS)

S5 Fig. PCA plots (PC1 vs. PC2) of the AMP-AD case-control dataset showing the distribution of ancestry backgrounds.
(EPS)

S6 Fig. PCA plots (PC1 vs. PC2) of the ADSP Family dataset showing the distribution of ancestry backgrounds.
(EPS)

S7 Fig. Overlapping genes between gene-sets (the endocytic, the immune response, and the lipid metabolism pathways) and the findings in recent GWASes. Gene-sets were defined through AmiGO 2 gene-ontology database. Two lists of genes implicated in AD were obtained from the two recent GWASes, Jansen et al.[1] (left) and Kunkle et al.[2] (right), and compared against the three defined gene-sets. The count of overlapping genes between each gene-set and

the findings from recent GWASes were shown above. To note, AD-implicated genes were identified through a variety of ways in the GWASes and the overlapping counts in each category were shown.
(DOCX)

S8 Fig. Distribution of CADD scores among rare deleterious variants defined by VEP and PolyPhen-2.
(DOCX)

S9 Fig. Distribution of rare deleterious variants in different mutation categories.
(DOCX)

S10 Fig. Distribution of pLI scores among endocytic genes.
(DOCX)

S11 Fig. Overlapping genes between the four gene-sets (endo-system, endosome, lysosome, and trans-Golgi network).
(DOCX)

S12 Fig. Comparison of age distribution between AD cases and controls in the three datasets (ADSP case-control, AMP-AD case-control, and ADSP Family datasets).
(DOCX)

S13 Fig. Functional annotation and confirmation of the biological functions of the endo-system gene-set.
(EPS)

S1 Table. Rare-variant gene-set AD association analysis using PLINK. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of APOE ε2 and ε4 alleles. The P and P* in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. # variants represented the number of rare deleterious variants identified in each dataset for each gene-set. The directions of effects were consistent across nearly all models.
(DOCX)

S2 Table. Comparison of stage 2 AMP-AD rare-variant gene-set AD association analysis. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The stage 2 AMP-AD* cohort represented the largest AMP-AD sub-cohort, ROSMAP.
(DOCX)

S3 Table. Rare-variant AD association analysis using the MAGMA burden method. The starred (*) geneset are those excluding the *APOE* gene. The Mu and P-self represented the estimated mean association and the self-contained p-value testing whether an association existed within the tested gene-set. The Beta and P-comp represented the estimated effect size and the

competitive p-value testing whether the association within the gene-set was greater than in other genes. P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The directions of effects were consistent across nearly all models.
(DOCX)

S4 Table. Rare-variant AD association analysis using the MAGMA SNP-wise method. The starred (*) geneset are those excluding the *APOE* gene. The Mu and P-self represented the estimated mean association and the self-contained p-value testing whether an association has existed within the tested gene-set. The Beta and P-comp represented the estimated effect size and the competitive p-value testing whether the association within the gene-set was greater than in other genes. P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The directions of effects were consistent across nearly all models.
(DOCX)

S5 Table. Rare-variant AD association analysis using PLINK where rare variants were annotated by a combination of VEP, PolyPhen-2, and CADD (>15). The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The stage 2 AMP-AD cohort was analyzed using all sub-cohorts and the largest ROSMAP sub-cohort (71.5% of the total sample size; marked in *). The P and P* in the meta-analysis across two stages (three datasets; AMP-AD* was used here) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. Similar results could be obtained using the stage 2 AMP-AD. The directions of effects were consistent across nearly all models.
(DOCX)

S6 Table. Rare-variant gene-set Braak association analysis using PLINK. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The P and P* in the meta-analysis across two stages (two datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. The directions of effects were consistent across nearly all models.
(DOCX)

S7 Table. Rare-variant gene-set AAO and AOD association analysis using PLINK. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The P and P* in the meta-analysis across two stages (two datasets) represented the p-values calculated

using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. The directions of effects were consistent across nearly all models.
(DOCX)

S8 Table. Top ten most significant genes in rare-variant single-gene NFT association analysis. The genes were sorted in the descending order of p-values. The meta-analysis was performed using MetaSKAT. P-values below the Bonferroni threshold ($\alpha = 4.83^{*}10^{-5}$; $4.25^{*}10^{-5}$; $5.17^{*}10^{-5}$, for ADSP, AMP-AD, and meta-analysis, respectively) were highlighted in red.
(DOCX)

S9 Table. Differential expression analysis of three identified genes, *HLA-A*, *SLC26A*, and *ANKRD13D*, between AD cases and controls from the ROSMAP study using six major cell types. Abbreviations: Ex: excitatory neuron; In: inhibitory neuron; Ast: astrocyte; Oli: oligodendrocyte; Opc: oligodendrocyte-precursor-cell; Mic: microglia. Effect: t-statistics calculated using student t-test, representing the direction of effect. P-values are computed using the same method.
(DOCX)

S10 Table. Differential expression analysis of three identified genes, *HLA-A*, *SLC26A*, and *ANKRD13D*, between AD cases and control from the ROSMAP study using four cellular subpopulations implicated with AD pathology. Abbreviations: Ex: excitatory neuron; In: inhibitory neuron; Ast: astrocyte; Oli: oligodendrocyte; Opc: oligodendrocyte-precursor-cell; Mic: microglia. Effect: t-statistics calculated using student t-test, representing the direction of effect. P-values are computed using the same method.
(DOCX)

S11 Table. Count of singletons and private doubletons within the included rare deleterious variants. The number of total variants represented all rare deleterious variants included under the corresponding MAF threshold.
(DOCX)

S12 Table. Rare-variant AD association analysis weighted by pLI scores. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The P and P^{*} in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively.
(DOCX)

S13 Table. Rare-variant AD association analysis using gene-sets related to BMI and height. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The P and P^{*} in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively.
(DOCX)

S14 Table. Number of rare variants passing different HWE cutoffs at different MAF thresholds. Abbreviations: HWE: Hardy-Weinberg Equilibrium; MAF: minor allele frequency.
(DOCX)

S15 Table. Top ten most significant genes in rare-variant single-gene AD association analysis. The genes were sorted in the descending order of p-values. The meta-analysis was performed using MetaSKAT. The Bonferroni thresholds were $\alpha = 4.18^{*}10^{-5}$; $4.07^{*}10^{-5}$; $7.32^{*}10^{-5}$; $7.79^{*}10^{-5}$, for ADSP, AMP-AD, ADSP Family, and meta-analysis, respectively.
(DOCX)

S16 Table. A full list of the endocytic genes with corresponding coordinates in GRCh38 and the average coverage (DP) in the ADSP and the AMP-AD datasets.
(XLSX)

Acknowledgments

We thank the ADSP, ADNI, and AMP-AD contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. The data used in this study can be found using accession NG00067 and Synapse portal. Specific acknowledgment to each study is as follows:

Acknowledgments to ADSP

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang).

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

The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the

neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904), the EU Joint Programme—Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN2682 01100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268 201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG 023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American

Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or non-governmental organizations.

Acknowledgments to ADNI

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Acknowledgments to AMP-AD

Mayo RNAseq Study- Study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24

NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05–901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research.

ROSMAP

We are grateful to the participants in the Religious Order Study, the Memory and Aging Project. This work is supported by the US National Institutes of Health [U01 AG046152, R01 AG043617, R01 AG042210, R01 AG036042, R01 AG036836, R01 AG032990, R01 AG18023, RC2 AG036547, P50 AG016574, U01 ES017155, KL2 RR024151, K25 AG041906-01, R01 AG30146, P30 AG10161, R01 AG17917, R01 AG15819, K08 AG034290, P30 AG10161 and R01 AG11101.

ROSMAP (gene expression data)

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu.

Mount Sinai Brain Bank (MSBB)

This work was supported by the grants R01AG046170, RF1AG054014, RF1AG057440 and R01AG057907 from the NIH/National Institute on Aging (NIA). R01AG046170 is a component of the AMP-AD Target Discovery and Preclinical Validation Project. Brain tissue collection and characterization was supported by NIH HHSN271201300031C.

Author Contributions

Conceptualization: Lingyu Zhan, Jae Hoon Sul.

Formal analysis: Lingyu Zhan.

Funding acquisition: Jae Hoon Sul.

Investigation: Lingyu Zhan.

Methodology: Lingyu Zhan, Jiajin Li, Jae Hoon Sul.

Project administration: Jae Hoon Sul.

Resources: Brandon Jew, Jae Hoon Sul.

Supervision: Jae Hoon Sul.

Validation: Lingyu Zhan.

Visualization: Lingyu Zhan.

Writing – original draft: Lingyu Zhan.

Writing – review & editing: Lingyu Zhan, Jae Hoon Sul.

References

1. Mendez MF. Early-onset Alzheimer's disease: nonamnestic subtypes and type 2 AD. Arch Med Res. 2012; 43(8):677–85. https://doi.org/10.1016/j.arcmed.2012.11.009 PMID: 23178565
2. Burns A, Iliffe S. Dementia. BMJ. 2009; 338:b75. https://doi.org/10.1136/bmj.b75 PMID: 19196746
3. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology. 2013; 80(19):1778–83. https://doi.org/10.1212/WNL.0b013e31828726f5 PMID: 23390181
4. 2021 Alzheimer's disease facts and figures. Alzheimers Dement. 2021; 17(3):327–406. https://doi.org/10.1002/alz.12328 PMID: 33756057
5. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007; 3(3):186–91. https://doi.org/10.1016/j.jalz.2007.04.381 PMID: 19595937
6. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011; 43(5):436–41. https://doi.org/10.1038/ng.801 PMID: 21460841
7. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006; 63(2):168–74. https://doi.org/10.1001/archpsyc.63.2.168 PMID: 16461860
8. Shen L, Jia J. An Overview of Genome-Wide Association Studies in Alzheimer's Disease. Neurosci Bull. 2016; 32(2):183–90. https://doi.org/10.1007/s12264-016-0011-3 PMID: 26810783
9. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013; 45(12):1452–8. https://doi.org/10.1038/ng.2802 PMID: 24162737
10. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics C. Alzheimer's disease: analyzing the missing heritability. PLoS One. 2013; 8(11):e79771. https://doi.org/10.1371/journal.pone.0079771 PMID: 24244562
11. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019; 51(3):404–13. https://doi.org/10.1038/s41588-018-0311-9 PMID: 30617256
12. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet. 2019; 51(3):414–30. https://doi.org/10.1038/s41588-019-0358-2 PMID: 30820047
13. Small SA, Simoes-Spassov S, Mayeux R, Petsko GA. Endosomal Traffic Jams Represent a Pathogenic Hub and Therapeutic Target in Alzheimer's Disease. Trends Neurosci. 2017; 40(10):592–602. https://doi.org/10.1016/j.tins.2017.08.003 PMID: 28962801
14. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013; 368(2):117–27. https://doi.org/10.1056/NEJMoa1211851 PMID: 23150934
15. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med. 2013; 368(2):107–16. https://doi.org/10.1056/NEJMoa1211103 PMID: 23150908
16. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature. 2014; 505(7484):550–4. https://doi.org/10.1038/nature12825 PMID: 24336208
17. Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, et al. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. Alzheimers Dement. 2014; 10(6):609–18 e11. https://doi.org/10.1016/j.jalz.2014.06.010 PMID: 25172201
18. Reitz C, Mayeux R, Alzheimer's Disease Genetics C. TREM2 and neurodegenerative disease. N Engl J Med. 2013; 369(16):1564–5. https://doi.org/10.1056/NEJMc1306509 PMID: 24131184

19. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017; 49(9):1373–84. https://doi.org/10.1038/ng.3916 PMID: 28714976

20. Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, et al. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. Nat Med. 2014; 20(12):1452–7. https://doi.org/10.1038/nm.3736 PMID: 25419706

21. Cataldo AM, Petanceska S, Terio NB, Peterhoff CM, Durham R, Mercken M, et al. Abeta localization in abnormal endosomes: association with earliest Abeta elevations in AD and Down syndrome. Neurobiol Aging. 2004; 25(10):1263–72. https://doi.org/10.1016/j.neurobiolaging.2004.02.027 PMID: 15465622

22. Cataldo A, Rebeck GW, Ghetri B, Hulette C, Lippa C, Van Broeckhoven C, et al. Endocytic disturbances distinguish among subtypes of Alzheimer's disease and related disorders. Ann Neurol. 2001; 50(5):661–5. PMID: 11706973

23. Corlier F, Rivals I, Lagarde J, Hamelin L, Corne H, Dauphinot L, et al. Modifications of the endosomal compartment in peripheral blood mononuclear cells and fibroblasts from Alzheimer's disease patients. Transl Psychiatry. 2015; 5:e595. https://doi.org/10.1038/tp.2015.87 PMID: 26151923

24. Cataldo AM, Peterhoff CM, Troncoso JC, Gomez-Isla T, Hyman BT, Nixon RA. Endocytic pathway abnormalities precede amyloid beta deposition in sporadic Alzheimer's disease and Down syndrome: differential effects of APOE genotype and presenilin mutations. Am J Pathol. 2000; 157(1):277–86. https://doi.org/10.1016/s0002-9440(10)64538-5 PMID: 10880397

25. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet. 2021; 53(3):392–402. https://doi.org/10.1038/s41588-020-00776-w PMID: 33589840

26. Heckmann BL, Teubner BJW, Tummers B, Boada-Romero E, Harris L, Yang M, et al. LC3-Associated Endocytosis Facilitates beta-Amyloid Clearance and Mitigates Neurodegeneration in Murine Alzheimer's Disease. Cell. 2020; 183(6):1733–4. https://doi.org/10.1016/j.cell.2020.11.033 PMID: 33306957

27. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiatry. 2015; 77(1):43–51. https://doi.org/10.1016/j.biopsych.2014.05.006 PMID: 24951455

28. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467(7317):832–8. https://doi.org/10.1038/nature09410 PMID: 20881960

29. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res. 2010; 70(11):4453–9. https://doi.org/10.1158/0008-5472.CAN-09-4502 PMID: 20460509

30. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015; 518(7538):197–206. https://doi.org/10.1038/nature14177 PMID: 25673413

31. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum Mol Genet. 2011; 20(17):3494–506. https://doi.org/10.1093/hmg/ddr248 PMID: 21653640

32. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet. 2009; 84(3):399–405. https://doi.org/10.1016/j.ajhg.2009.01.026 PMID: 19249008

33. Nurnberger JI Jr., Koller DL, Jung J, Edenberg HJ, Foroud T, Guella I, et al. Identification of pathways for bipolar disorder: a meta-analysis. JAMA Psychiatry. 2014; 71(6):657–64. https://doi.org/10.1001/jamapsychiatry.2014.176 PMID: 24718920

34. Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet. 2009; 125(1):63–79. https://doi.org/10.1007/s00439-008-0600-y PMID: 19052778

35. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421–7. https://doi.org/10.1038/nature13595 PMID: 25056061

36. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landen M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016; 19(11):1433–41. https://doi.org/10.1038/nn.4402 PMID: 27694994

37. Manshaei R, Merico D, Reuter MS, Engchuan W, Mojarad BA, Chaturvedi R, et al. Genes and Pathways Implicated in Tetralogy of Fallot Revealed by Ultra-Rare Variant Burden Analysis in 231 Genome Sequences. Front Genet. 2020; 11:957. https://doi.org/10.3389/fgene.2020.00957 PMID: 33110418

38. Amanat S, Gallego-Martinez A, Sollini J, Perez-Carpena P, Espinosa-Sanchez JM, Aran I, et al. Burden of rare variants in synaptic genes in patients with severe tinnitus: An exome based extreme phenotype study. EBioMedicine. 2021; 66:103309. https://doi.org/10.1016/j.ebiom.2021.103309 PMID: 33813136

39. Sul JH, Service SK, Huang AY, Ramensky V, Hwang SG, Teshiba TM, et al. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. Transl Psychiatry. 2020; 10(1):74. https://doi.org/10.1038/s41398-020-0758-1 PMID: 32094344

40. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013; 45(10):1150–9. https://doi.org/10.1038/ng.2742 PMID: 23974872

41. Qian DC, Byun J, Han Y, Greene CS, Field JK, Hung RJ, et al. Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions. Hum Mol Genet. 2015; 24(25):7406–20. https://doi.org/10.1093/hmg/ddv440 PMID: 26483192

42. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 2013; 14 (8):549–58. https://doi.org/10.1038/nrg3523 PMID: 23835440

43. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet. 2012; 13(8):537–51. https://doi.org/10.1038/nrg3240 PMID: 22777127

44. Schizophrenia Psychiatric Genome-Wide Association Study C. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011; 43(10):969–76. https://doi.org/10.1038/ng.940 PMID: 21926974

45. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363(2):166–76. https://doi.org/10.1056/NEJMra0905980 PMID: 20647212

46. Xiao X, Jiao B, Liao X, Zhang W, Yuan Z, Guo L, et al. Association of Genes Involved in the Metabolic Pathways of Amyloid-beta and Tau Proteins With Sporadic Late-Onset Alzheimer's Disease in the Southern Han Chinese Population. Front Aging Neurosci. 2020; 12:584801. https://doi.org/10.3389/fnagi.2020.584801 PMID: 33240075

47. Xiang S, Huang Z, Wang T, Han Z, Yu CY, Ni D, et al. Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. BMC Med Genomics. 2018; 11(Suppl 6):115. https://doi.org/10.1186/s12920-018-0431-1 PMID: 30598117

48. Abner EL, Kryscio RJ, Schmitt FA, Santacruz KS, Jicha GA, Lin Y, et al. "End-stage" neurofibrillary tangle pathology in preclinical Alzheimer's disease: fact or fiction? J Alzheimers Dis. 2011; 25(3):445–53. https://doi.org/10.3233/JAD-2011-101980 PMID: 21471646

49. Braak H, Braak E, Bohl J. Staging of Alzheimer-related cortical destruction. Eur Neurol. 1993; 33 (6):403–8. https://doi.org/10.1159/000116984 PMID: 8307060

50. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991; 82(4):239–59. https://doi.org/10.1007/BF00308809 PMID: 1759558

51. Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, et al. The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet. 2017; 3(5):e194. https://doi.org/10.1212/NXG.0000000000000194 PMID: 29184913

52. ADSP Discovery Extension Case-Control Sample Selection Criteria.

53. Leung YY, Valladares O, Chou YF, Lin HJ, Kuzma AB, Cantwell L, et al. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. Bioinformatics. 2019; 35(10):1768–70. https://doi.org/10.1093/bioinformatics/bty894 PMID: 30351394

54. ADNI procedue manual online protocol.

55. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018; 5:180142. https://doi.org/10.1038/sdata.2018.142 PMID: 30084846

56. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012; 9(6):628–45. https://doi.org/10.2174/156720512801322573 PMID: 22471860

57. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016; 3:160089. https://doi.org/10.1038/sdata.2016.89 PMID: 27727239

58. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018; 5:180185. https://doi.org/10.1038/sdata.2018.185 PMID: 30204156

59. Blue EE, Bis JC, Dorschner MO, Tsuang DW, Barral SM, Beecham G, et al. Genetic Variation in Genes Underlying Diverse Dementias May Explain a Small Proportion of Cases in the Alzheimer's Disease Sequencing Project. Dement Geriatr Cogn Disord. 2018; 45(1–2):1–17. https://doi.org/10.1159/000485503 PMID: 29486463

60. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2020; 25(8):1859–75. https://doi.org/10.1038/s41380-018-0112-7 PMID: 30108311

61. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019; 570(7761):332–7. https://doi.org/10.1038/s41586-019-1195-2 PMID: 31042697

62. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81 (3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

63. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

64. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–9. https://doi.org/10.1038/ng1847 PMID: 16862161

65. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

66. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):122. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

67. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7(4):248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512

68. Saint Pierre A, Genin E. How important are rare variants in common disease? Brief Funct Genomics. 2014; 13(5):353–61. https://doi.org/10.1093/bfgp/elu025 PMID: 25005607

69. Todorovic V. Genetics. Predicting the impact of genomic variation. Nat Methods. 2016; 13(3):203. https://doi.org/10.1038/nmeth.3793 PMID: 27347591

70. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021; 13(1):31. https://doi.org/10.1186/s13073-021-00835-9 PMID: 33618777

71. Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021; 49 (D1):D325–D34. https://doi.org/10.1093/nar/gkaa1113 PMID: 33290552

72. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–9. https://doi.org/10.1038/75556 PMID: 10802651

73. Mellman I. Endocytosis and molecular sorting. Annu Rev Cell Dev Biol. 1996; 12:575–625. https://doi.org/10.1146/annurev.cellbio.12.1.575 PMID: 8970738

74. Nakano A, Luini A. Passage through the Golgi. Curr Opin Cell Biol. 2010; 22(4):471–8. https://doi.org/10.1016/j.ceb.2010.05.003 PMID: 20605430

75. Settembre C, Fraldi A, Medina DL, Ballabio A. Signals from the lysosome: a control centre for cellular clearance and energy metabolism. Nat Rev Mol Cell Biol. 2013; 14(5):283–96. https://doi.org/10.1038/nrm3565 PMID: 23609508

76. Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics. 2012; 28(13):1805–6. https://doi.org/10.1093/bioinformatics/bts251 PMID: 22543366

77. Zhang X, Zhu C, Beecham G, Vardarajan BN, Ma Y, Lancour D, et al. A rare missense variant of CASP7 is associated with familial late-onset Alzheimer's disease. Alzheimers Dement. 2019; 15 (3):441–52. https://doi.org/10.1016/j.jalz.2018.10.005 PMID: 30503768

78. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016; 98(4):653–66. https://doi.org/10.1016/j.ajhg.2016.02.012 PMID: 27018471

79. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015; 11(4):e1004219. https://doi.org/10.1371/journal.pcbi.1004219 PMID: 25885710

80. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26(17):2190–1. https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382

81. Chen Z, Yang W, Liu Q, Yang JY, Li J, Yang M. A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. BMC Bioinformatics. 2014; 15 Suppl 17:S3. https://doi.org/10.1186/1471-2105-15-S17-S3 PMID: 25559433

82. Chen Z, Huang H, Liu J, Tony Ng HK, Nadarajah S, Huang X, et al. Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data. BMC Med Genomics. 2013; 6 Suppl 1:S9. https://doi.org/10.1186/1755-8794-6-S1-S9 PMID: 23369576

83. Dewey M. metap: meta-analysis of significance values. 2020.

84. Ripley WNVaBD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.

85. Gotzsche PC. Why we need a broad perspective on meta-analysis. It may be crucially important for patients. BMJ. 2000; 321(7261):585–6. https://doi.org/10.1136/bmj.321.7261.585 PMID: 10977820

86. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 2017; 14(3):e1002258. https://doi.org/10.1371/journal.pmed.1002258 PMID: 28323831

87. Terry M. Therneau PMG. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

88. A Kassambara MK P Biecek. survminer: Drawing Survival Curves using 'ggplot2'. 2017.

89. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059

90. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013; 93(1):42–53. https://doi.org/10.1016/j.ajhg.2013.05.010 PMID: 23768515

91. Sul JH, Raj T, de Jong S, de Bakker PI, Raychaudhuri S, Ophoff RA, et al. Accurate and fast multiple-testing correction in eQTL studies. Am J Hum Genet. 2015; 96(6):857–68. https://doi.org/10.1016/j.ajhg.2015.04.012 PMID: 26027500

92. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013; 9(5):e1003486. https://doi.org/10.1371/journal.pgen.1003486 PMID: 23671422

93. Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci. 2019; 22(12):2087–97. https://doi.org/10.1038/s41593-019-0539-4 PMID: 31768052

94. De Strooper B, Karran E. The Cellular Phase of Alzheimer's Disease. Cell. 2016; 164(4):603–15. https://doi.org/10.1016/j.cell.2015.12.056 PMID: 26871627

95. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018; 36(5):411–20. https://doi.org/10.1038/nbt.4096 PMID: 29608179

96. Hu YB, Dammer EB, Ren RJ, Wang G. The endosomal-lysosomal system: from acidification and cargo sorting to neurodegeneration. Transl Neurodegener. 2015; 4:18. https://doi.org/10.1186/s40035-015-0041-1 PMID: 26448863

97. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016; 2016. https://doi.org/10.1093/database/baw100 PMID: 27374120

98. Iwakiri M, Mizukami K, Ikonomovic MD, Ishikawa M, Hidaka S, Abrahamson EE, et al. Changes in hippocampal GABABR1 subunit expression in Alzheimer's patients: association with Braak staging. Acta Neuropathol. 2005; 109(5):467–74. https://doi.org/10.1007/s00401-005-0985-9 PMID: 15759131

99. Singh V, Dwivedi SN, Deo SVS. Ordinal logistic regression model describing factors associated with extent of nodal involvement in oral cancer patients and its prospective validation. BMC Med Res Methodol. 2020; 20(1):95. https://doi.org/10.1186/s12874-020-00985-1 PMID: 32336269

100. Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet. 2014; 10(9):e1004606. https://doi.org/10.1371/journal.pgen.1004606 PMID: 25188341

101. Li YJ, Scott WK, Hedges DJ, Zhang F, Gaskell PC, Nance MA, et al. Age at onset in two common neurodegenerative diseases is genetically controlled. Am J Hum Genet. 2002; 70(4):985–93. https://doi.org/10.1086/339815 PMID: 11875758

102. Daw EW, Payami H, Nemens EJ, Nochlin D, Bird TD, Schellenberg GD, et al. The number of trait loci in late-onset Alzheimer disease. Am J Hum Genet. 2000; 66(1):196–204. https://doi.org/10.1086/302710 PMID: 10631151

103. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018; 8(1):99. https://doi.org/10.1038/s41398-018-0150-6 PMID: 29777097

104. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. Nat Neurosci. 2017; 20(8):1052–61. https://doi.org/10.1038/nn.4587 PMID: 28628103

105. Naj AC, Jun G, Reitz C, Kunkle BW, Perry W, Park YS, et al. Effects of multiple genetic loci on age at onset in late-onset Alzheimer disease: a genome-wide association study. JAMA Neurol. 2014; 71(11):1394–404. https://doi.org/10.1001/jamaneurol.2014.1491 PMID: 25199842

106. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. Nat Genet. 2017; 49(3):325–31. https://doi.org/10.1038/ng.3766 PMID: 28092683

107. Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Mol Psychiatry. 2012; 17(12):1340–6. https://doi.org/10.1038/mp.2011.135 PMID: 22005931

108. Zhang Q, Sidorenko J, Couvy-Duchesne B, Marioni RE, Wright MJ, Goate AM, et al. Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun. 2020; 11(1):4799. https://doi.org/10.1038/s41467-020-18534-1 PMID: 32968074

109. Moore KM, Nicholas J, Grossman M, McMillan CT, Irwin DJ, Massimo L, et al. Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study. Lancet Neurol. 2020; 19(2):145–56. https://doi.org/10.1016/S1474-4422(19)30394-1 PMID: 31810826

110. Shi H, Belbin O, Medway C, Brown K, Kalsheker N, Carrasquillo M, et al. Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS). Neurobiol Aging. 2012; 33(8):1849 e5-18. https://doi.org/10.1016/j.neurobiolaging.2012.02.014 PMID: 22445811

111. Bai Z, Stamova B, Xu H, Ander BP, Wang J, Jickling GC, et al. Distinctive RNA expression profiles in blood associated with Alzheimer disease after accounting for white matter hyperintensities. Alzheimer Dis Assoc Disord. 2014; 28(3):226–33. https://doi.org/10.1097/WAD.0000000000000022 PMID: 24731980

112. Haenig C, Atias N, Taylor AK, Mazza A, Schaefer MH, Russ J, et al. Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains. Cell Rep. 2020; 32(7):108050. https://doi.org/10.1016/j.celrep.2020.108050 PMID: 32814053

113. Yu L, Chibnik LB, Srivastava GP, Pochet N, Yang J, Xu J, et al. Association of Brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease. JAMA Neurol. 2015; 72(1):15–24. https://doi.org/10.1001/jamaneurol.2014.3049 PMID: 25365775

114. Vico Varela E, Etter G, Williams S. Excitatory-inhibitory imbalance in Alzheimer's disease and therapeutic significance. Neurobiol Dis. 2019; 127:605–15. https://doi.org/10.1016/j.nbd.2019.04.010 PMID: 30999010

115. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nat Rev Neurol. 2021; 17(3):157–72. https://doi.org/10.1038/s41582-020-00435-y PMID: 33318676

116. Van Acker ZP, Bretou M, Annaert W. Endo-lysosomal dysregulations and late-onset Alzheimer's disease: impact of genetic risk factors. Mol Neurodegener. 2019; 14(1):20. https://doi.org/10.1186/s13024-019-0323-7 PMID: 31159836

117. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry. 2006; 45(51):15168–78. https://doi.org/10.1021/bi062188q PMID: 17176038

118. Tanno H, Yamaguchi T, Goto E, Ishido S, Komada M. The Ankrd 13 family of UIM-bearing proteins regulates EGF receptor endocytosis from the plasma membrane. Mol Biol Cell. 2012; 23(7):1343–53. https://doi.org/10.1091/mbc.E11-09-0817 PMID: 22298428

119. Burana D, Yoshihara H, Tanno H, Yamamoto A, Saeki Y, Tanaka K, et al. The Ankrd13 Family of Ubiquitin-interacting Motif-bearing Proteins Regulates Valosin-containing Protein/p97 Protein-mediated Lysosomal Trafficking of Caveolin 1. J Biol Chem. 2016; 291(12):6218–31. https://doi.org/10.1074/jbc.M115.710707 PMID: 26797118

120. Humpel C, Hochstrasser T. Cerebrospinal fluid and blood biomarkers in Alzheimer's disease. World J Psychiatry. 2011; 1(1):8–18. https://doi.org/10.5498/wjp.v1.i1.8 PMID: 24175162

121. Doecke JD, Laws SM, Faux NG, Wilson W, Burnham SC, Lam CP, et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. Arch Neurol. 2012; 69(10):1318–25. https://doi.org/10.1001/archneurol.2012.1282 PMID: 22801742

122. Bjorkqvist M, Ohlsson M, Minthon L, Hansson O. Evaluation of a previously suggested plasma biomarker panel to identify Alzheimer's disease. PLoS One. 2012; 7(1):e29868. https://doi.org/10.1371/journal.pone.0029868 PMID: 22279551

123. Thomas R, Zuchowska P, Morris AW, Marottoli FM, Sunny S, Deaton R, et al. Epidermal growth factor prevents APOE4 and amyloid-beta-induced cognitive and cerebrovascular deficits in female mice. Acta Neuropathol Commun. 2016; 4(1):111. https://doi.org/10.1186/s40478-016-0387-3 PMID: 27788676

124. Koster KP, Thomas R, Morris AWJ, Tai LM. Epidermal growth factor prevents oligomeric amyloid-beta induced angiogenesis deficits in vitro. J Cereb Blood Flow Metab. 2016; 36(11):1865–71. https://doi.org/10.1177/0271678X16669956 PMID: 27634936

125. Cresswell P, Ackerman AL, Giodini A, Peaper DR, Wearsch PA. Mechanisms of MHC class I-restricted antigen processing and cross-presentation. Immunol Rev. 2005; 207:145–57. https://doi.org/10.1111/j.0105-2896.2005.00316.x PMID: 16181333

126. Basha G, Lizee G, Reinicke AT, Seipp RP, Omilusik KD, Jefferies WA. MHC class I endosomal and lysosomal trafficking coincides with exogenous antigen loading in dendritic cells. PLoS One. 2008; 3(9):e3247. https://doi.org/10.1371/journal.pone.0003247 PMID: 18802471

127. Lazarczyk MJ, Kemmler JE, Eyford BA, Short JA, Varghese M, Sowa A, et al. Major Histocompatibility Complex class I proteins are critical for maintaining neuronal structural complexity in the aging brain. Sci Rep. 2016; 6:26199. https://doi.org/10.1038/srep26199 PMID: 27229916

128. Ma SL, Tang NL, Tam CW, Lui VW, Suen EW, Chiu HF, et al. Association between HLA-A alleles and Alzheimer's disease in a southern Chinese community. Dement Geriatr Cogn Disord. 2008; 26(5):391–7. https://doi.org/10.1159/000164275 PMID: 18936542

129. Guerini FR, Tinelli C, Calabrese E, Agliardi C, Zanzottera M, De Silvestri A, et al. HLA-A*01 is associated with late onset of Alzheimer's disease in Italian patients. Int J Immunopathol Pharmacol. 2009; 22(4):991–9. https://doi.org/10.1177/039463200902200414 PMID: 20074462

130. Liu Y, Liu F, Iqbal K, Grundke-Iqbal I, Gong CX. Decreased glucose transporters correlate to abnormal hyperphosphorylation of tau in Alzheimer disease. FEBS Lett. 2008; 582(2):359–64. https://doi.org/10.1016/j.febslet.2007.12.035 PMID: 18174027

131. Alper SL, Sharma AK. The SLC26 gene family of anion transporters and channels. Mol Aspects Med. 2013; 34(2–3):494–515. https://doi.org/10.1016/j.mam.2012.07.009 PMID: 23506885

132. Yin K, Lei Y, Wen X, Lacruz RS, Soleimani M, Kurtz I, et al. SLC26A Gene Family Participate in pH Regulation during Enamel Maturation. PLoS One. 2015; 10(12):e0144703. https://doi.org/10.1371/journal.pone.0144703 PMID: 26671068

133. Hooijmans CR, Graven C, Dederen PJ, Tanila H, van Groen T, Kiliaan AJ. Amyloid beta deposition is related to decreased glucose transporter-1 levels and hippocampal atrophy in brains of aged APP/PS1 mice. Brain Res. 2007; 1181:93–103. https://doi.org/10.1016/j.brainres.2007.08.063 PMID: 17916337

134. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016; 48(2):214–20. https://doi.org/10.1038/ng.3477 PMID: 26727659

135. Holstege H, Hulsman M, Charbonnier C, Grenier-Boley B, Quenez O, Grozeva D, et al. Exome sequencing identifies novel AD-associated genes. medRxiv. 2020:2020.07.22.20159251.

136. NIAGADS D. DSS Release Notes—NG00067.v6. 2021.