

RESEARCH

Open Access



# Genome-wide epistasis analysis for Alzheimer's disease and implications for genetic risk prediction

Hui Wang<sup>1</sup>, David A. Bennett<sup>2,3</sup>, Philip L. De Jager<sup>4,5</sup>, Qing-Ye Zhang<sup>1</sup> and Hong-Yu Zhang<sup>1\*</sup> 

## Abstract

**Background:** Single-nucleotide polymorphisms (SNPs) identified by genome-wide association studies only explain part of the heritability of Alzheimer's disease (AD). Epistasis has been considered as one of the main causes of "missing heritability" in AD.

**Methods:** We performed genome-wide epistasis screening ( $N = 10,389$ ) for the clinical diagnosis of AD using three popularly adopted methods. Subsequent analyses were performed to eliminate spurious associations caused by possible confounding factors. Then, candidate genetic interactions were examined for their co-expression in the brains of AD patients and analyzed for their association with intermediate AD phenotypes. Moreover, a new approach was developed to compile the epistasis risk factors into an epistasis risk score (ERS) based on multifactor dimensional reduction. Two independent datasets were used to evaluate the feasibility of ERSs in AD risk prediction.

**Results:** We identified 2 candidate genetic interactions with  $P_{FDR} < 0.05$  (*RAMP3-SEMA3A* and *NSMCE1-DGKE/C17orf67*) and another 5 genetic interactions with  $P_{FDR} < 0.1$ . Co-expression between the identified interactions supported the existence of possible biological interactions underlying the observed statistical significance. Further association of candidate interactions with intermediate phenotypes helps explain the mechanisms of neuropathological alterations involved in AD. Importantly, we found that ERSs can identify high-risk individuals showing earlier onset of AD. Combined risk scores of SNPs and SNP-SNP interactions showed slightly but steadily increased AUC in predicting the clinical status of AD.

**Conclusions:** In summary, we performed a genome-wide epistasis analysis to identify novel genetic interactions potentially implicated in AD. We found that ERS can serve as an indicator of the genetic risk of AD.

**Keywords:** Alzheimer's disease, Association studies in genetics, Gene expression studies, Polygenic risk score

\* Correspondence: [zhy630@mail.hzau.edu.cn](mailto:zhy630@mail.hzau.edu.cn)

<sup>1</sup>Huazhong Agricultural University, College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, No.1 Shizishan Street, Hongshan District, Wuhan 430070, Hubei, China

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

Alzheimer's disease (AD) is a chronic neurodegenerative disease that is characterized by the extracellular deposition of beta-amyloid and the intracellular accumulation of phosphorylated tau protein. AD is the most common cause of dementia in elderly people, with an incidence rate of approximately 1.5% among people over 65 years old and nearly 50% among people over 90 years old [1]. Unlike early-onset AD, which is often caused by mutations in *APP*, *PSEN1*, or *PSEN2* [2], late-onset AD (LOAD), the most common form of AD, exhibits a more complex genetic mechanism. Apolipoprotein  $\epsilon 4$  allele (*APOE4*) is the only common high-risk genetic variant associated with LOAD, and previous large-scale genome-wide association studies (GWASs) have identified dozens of loci with small effects [3–5], suggesting that a large portion of the genetic components of LOAD remains unexplained.

It is estimated that 24–33% of the phenotypic variance of LOAD can be explained by *APOE* combined with common variants [6, 7]; this value is considerably lower than the well-reported 58–79% of heritability estimated from twin studies [8]. Rare variants, structural variants, and genetic interactions are possible causes of missing heritability in complex diseases [9]. Previous studies have identified rare coding variants in *SORL1* and *ABCA7*, which can affect APP processing [10, 11]. Besides, rare variants in *PLCG2*, *ABI3*, and *TREM2* revealed the involvement of microglial-mediated innate immunity in AD [12]. In this study, we focus on the genetic interactions, which refers to the combinatorial effect of one or more variants, to help explain the missing heritability in AD. However, there are several challenges in detecting genetic interactions on a genome-wide scale. First, the computational burden of testing pairwise interactions exhaustively is heavy due to the quadratic complexity involved [13]. While a typical GWAS analysis analyzes several million SNPs, corresponding genome-wide interaction screening needs to be performed on more than  $1 \times 10^{14}$  SNP interactions, which is a prohibitive number. Second, the detection of genetic interactions is a typical case of a “large p, small n” problem [14]. To reduce the high rate of false positives caused by the astronomical number of tests performed, *P* value thresholds tend to be extremely conservative, while the sample size is usually the same as in traditional GWAS analysis, which can lead to a failure in discovering significant genetic interactions. Finally, the biological interpretation of statistical interactions is challenging, as statistical interactions do not necessarily imply an interaction at the biological level [15]. This situation is further complicated by the problem of insufficient sample size, as samples are stratified into the 9 cells of a  $3 \times 3$  contingency table instead of the 3 groups discriminated by the counts of

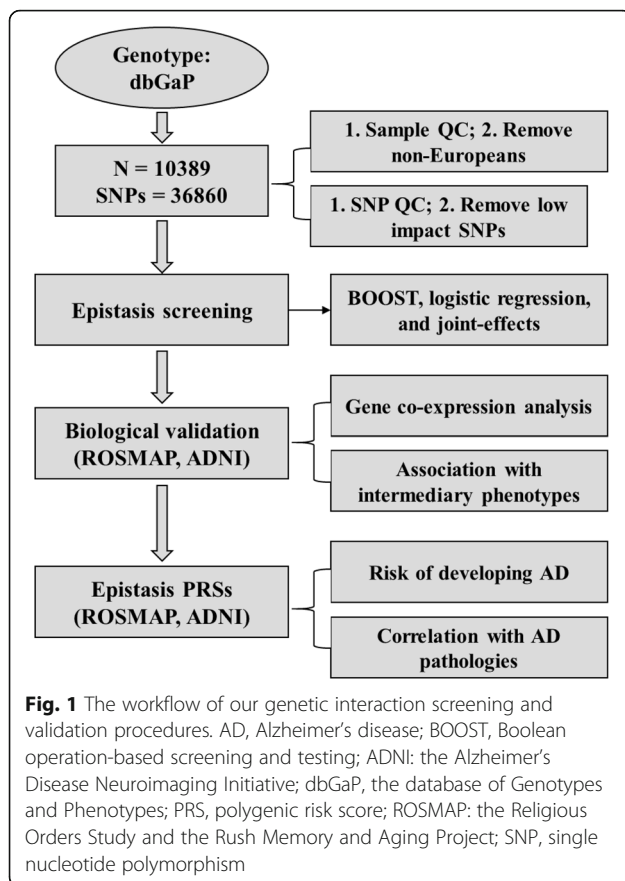
minor alleles in a typical GWAS analysis. The small sample size in the cells of the  $3 \times 3$  contingency table could lead to invalid biological interpretations of statistical interactions. Due to these limitations, only one genome-wide interaction analysis has identified an interaction between rs6455128 (*KHDRBS2*) and rs7989332 (*CRYL1*) that is replicable across datasets [16]. In this study, we limited the analysis to SNPs that are more likely to be deleterious according to combined annotation-dependent depletion (CADD) scoring. Therefore, the number of tests to be performed was scaled down. The aforementioned first and second problems of genetic interaction screening were alleviated. For better biological interpretation of statistical interactions, we excluded interactions with any of the cells containing less than 3 samples in the  $3 \times 3 \times 2$  contingency table. Furthermore, we analyzed the associations of candidate interactions with intermediate AD pathologies, including brain atrophy, white matter injury, and tau and amyloid deposition.

The polygenic architecture of AD enables the construction of predictive models based on genome-wide significant polymorphisms. A previous analysis of polygenic risk scores (PRSs) based on 22 GWAS-identified SNPs showed that PRSs were associated with the risk of AD and cerebrospinal fluid  $\beta$ -amyloid (1-42) ( $A\beta 1-42$ ) [17]. Another study showed that elevated PRSs were associated with worse memory and a smaller hippocampus at baseline, as well as greater longitudinal cognitive decline and clinical progression [18, 19]. In these analyses, PRSs displayed a significant but only moderate association with the trait of interest. Thus, PRSs alone seemed to be insufficient to capture the whole genetic landscape of AD. Through identifying genetic interactions, we attempted to evaluate the predictive capacities of epistasis risk scores as a complement to traditional PRSs. The whole workflow of our proposed analysis is displayed in Fig. 1.

## Methods

### Study subjects

Three independent datasets were used for genome-wide epistasis screening. National Institute on Aging (NIA)-LOAD (dbGaP accession: phs000168.v2.p2) study was to identify and recruit families with two or more siblings with the late-onset form of AD and a cohort of unrelated, non-demented controls similar in age and ethnic background, and to make the samples, the clinical and genotyping data and preliminary analyses available to qualified investigators worldwide [20]. Genetic Alzheimer's Disease Associations (GenADA, dbGaP accession: phs000219.v1.p1) was a multi-site collaborative study to associate DNA sequence (allelic) variations in candidate genes with AD [21]. The NIA Alzheimer's Disease Centers (ADCs, dbGaP accession: phs000372.v1.p1) cohort



consisted of autopsy-confirmed and clinically confirmed AD cases, cognitively normal elders (CNEs) with complete neuropathology data who were older than 60 years of age at death and living CNEs who were documented to not exhibit mild cognitive impairment (MCI) and were between 60 and 100 years of age at assessment [22].

Two additional independent datasets were used for the construction and testing of PRSs. The Religious Orders Study and the Rush Memory and Aging Project (ROS/MAP) study were longitudinal clinical-pathologic cohort studies of AD [23]. The diagnosis of AD for each subject was performed by a neurologist who reviewed all available clinical data each year blinded to prior years and at the time of death for all years blinded to all pathologic data. The Alzheimer's Disease Neuroimaging Initiative (ADNI, including phases 1, GO, and 2) was an international cooperative study conducted to investigate the pathology of AD and to develop treatments to slow or stop AD progression [24]. All subjects were administered clinical evaluations at the time of enrollment by expert physicians.

#### Genotyping, imputation, and sample quality control

The NIA-LOAD samples were genotyped using the Illumina Human610-Quad BeadChip (Illumina, Inc., San

Diego, CA, USA). The GenADA samples were genotyped using the Human Mapping 500 K Array Set (Affymetrix, Inc., Santa Clara, CA, USA). Samples from the ADCs were genotyped in two batches by using either the Human660W-Quad BeadChip or the HumanOmniExpress BeadChip (Illumina Inc., San Diego, CA, USA). There were 11,896 samples with both genotyping data and a clinical diagnosis of AD available in the three datasets.

Individuals in the ROS/MAP cohort were genotyped in two batches with a total sample size of 2090. The first batch was genotyped using the Affymetrix GeneChip 6.0 (Affymetrix, Inc., Santa Clara, CA, USA) at the Broad Institute's Center for Genotyping or the Translational Genomics Research Institute. The other batch was genotyped using the Illumina HumanOmniExpress platform (Illumina, Inc., San Diego, CA, USA) at the Children's Hospital of Philadelphia. A total of 1550 subjects from ADNI were genotyped with two platforms. A total of 757 individuals in ADNI1 were genotyped using the Illumina Human610-Quad BeadChip (Illumina, Inc., San Diego, CA, USA). A total of 793 ADNI/GO/2 subjects were genotyped using the HumanOmniExpress BeadChip (Illumina Inc., San Diego, CA, USA).

All datasets were phased using Eagle (v2.4.1) [25] and imputed using Minimac3 [26]. Genotyping data were first aligned to the human GRCh37/hg19 assembly using UCSC's liftOver tool [27]. Then, allele filtering and imputation were carried out as described in a previous study [28] with 1000 Genomes Phase3 integrated haplotypes as the reference panel [29]. Imputed variants with an imputation quality statistic ( $R^2$ ) below 0.3 were discarded.

Only individuals of European descent were selected for further analysis using GRAF-pop [30]. We excluded samples with a genotype missing rate  $> 0.2$  or heterozygosity rate  $\pm 3$  standard deviations from the mean. We removed individuals with discordant sex information between the input dataset and those imputed from X chromosome inbreeding coefficients. Then, samples from NIA-LOAD, GenADA, and ADCs were merged into one large dataset ( $N = 10,389$ ). Two batches from ROS/MAP were merged into one dataset ( $N = 2079$ ). The ADNI1/GO/2 data were merged into one dataset, which will hereafter be referred to as the ADNI dataset ( $N = 1419$ ). After sample quality control, 10,389 individuals were kept in the discovery dataset consisted of NIA-LOAD, GenADA, and ADCs (Table 1). For ROS/MAP, of 2090 individuals, 2079 were kept for further analysis (Table 2). For ADNI, of 1550 individuals, 1419 were kept for further analysis (Table 2).

#### SNP selection and quality control

We selected SNPs that were more likely to be deleterious based on combined annotation-dependent depletion

**Table 1** Characteristics of study participants ( $N = 10,389$ ) after QC of genetic data

Cohort	AD			Non-AD		
	N	Age, years (SD)	Sex (F/M)	N	Age, years (SD)	Sex (F/M)
NIA-LOAD	1832	76.14 (7.10)	650 M, 1182 F	1986	76.38 (8.55)	790 M, 1196 F
GenADA	805	78.04 (8.60)	340 M, 465 F	779	73.41 (7.92)	278 M, 501 F
ADCs	3516	72.00 (9.27)	1655 M, 1861 F	1471	75.90 (9.56)	542 M, 929 F
<b>Total</b>	<b>6153</b>	<b>74.02 (8.92)</b>	<b>2645 M, 3508 F</b>	<b>4236</b>	<b>75.43 (8.96)</b>	<b>1610 M, 2626 F</b>

NIA-LOAD, National Institute of Aging-Late Onset Alzheimer's Disease; GenADA, Genetic Alzheimer's Disease Associations; ADCs, Alzheimer's Disease Centers; AD, Alzheimer's disease; F, female; M, male; SD, standard deviation

(CADD) scores [31]. After imputation, only SNPs that were located within 5 kb of any protein-coding gene with a CADD score  $\geq 15$  were retained for further analysis. Furthermore, calls with an uncertainty greater than 0.2 or import dosage certainty smaller than 0.8 were treated as missing using PLINK (v1.90b4.10) [32]. Then, SNPs with a missing rate  $> 0.1$ , minor allele frequency  $< 0.05$ , or Hardy-Weinberg equilibrium test value of  $P < 1 \times 10^{-6}$  were removed. Ultimately, 36,860 SNPs passed the filtering and quality control processes.

#### AD pathologies and neuroimaging

Intermediate phenotypes such as AD pathologies and neuroimaging data can help understand how the identified genetic interactions work. Subsets of samples from ROS/MAP and ADNI have AD pathologies and neuroimaging data available. In ROS/MAP, immunohistochemistry and automated image processing were used for the measurement of total amyloid and paired helical filament tau (PHF-tau). A modified Bielschowsky silver staining technique was used to measure neuritic plaques, diffuse plaques, and neurofibrillary tangles. Among 2079 subjects in ROS/MAP, at the time of these analyses, 1310 had available measurements of neurofibrillary tangles, neuritic plaques, and diffuse plaques; 1279 had available total PHF-tau measurements; and 1270 had

available total amyloid measurements. In ADNI, cerebrospinal fluid (CSF) total tau (T-tau), phosphorylated tau (P-tau), and  $\beta$ -amyloid (1-42) ( $A\beta_{1-42}$ ) levels were measured via electrochemiluminescence immunoassays. Among 1419 individuals in the ADNI cohort, 1043 had available CSF T-tau, CSF P-tau, and CSF  $A\beta_{1-42}$  measurements. Molecular and structural neuroimaging data were also available for a subset of ADNI individuals. Structural magnetic resonance imaging was employed to generate estimates of the entorhinal cortex and hippocampal volume for 782 individuals. Fractional anisotropy (FA) for five bilateral fronto-temporal-occipital and interhemispheric white matter tracts (the sagittal stratum, hippocampal segment of the cingulum bundle, splenium of the corpus callosum, inferior fronto-occipital fasciculus, and superior longitudinal fasciculus) was estimated from diffusion-weighted images for 188 subjects.

#### Association and interaction analysis

Association analysis was used to evaluate the main effects of selected SNPs. A linear mixed model [33] was employed to detect the association between SNPs and the AD status with sex, age, imputation batch, and the first three principal components as covariates. The discovery dataset consisted of data from three studies (NIA-LOAD, GenADA, and ADCs) in four batches. To

**Table 2** Characteristics of study participants from ROS/MAP ( $N = 2079$ ) and ADNI1/GO/2 ( $N = 1419$ ) after QC of genetic data

	AD ( $n = 564$ )	Non-AD <sup>a</sup> ( $n = 1515$ )	Diff (P) <sup>b</sup>
<b>ROS/MAP</b>			
Sex (F/M)	388 F, 176 M	1060 F, 455 M	0.63
Age at death, years (SD)	90.92 (5.88)	88.10 (6.73)	$< 0.0001$
<i>APOE</i> $\epsilon 4$ status (-/+)	- 378, + 186 (0.33)	- 1226, + 289 (0.19)	$< 0.0001$
<b>ADNI</b>			
Sex (F/M)	231 F, 324 M	372 F, 492 M	0.62
Age at AD, years (SD)	74.78 (8.11)	77.49 (7.33)	$< 0.0001$
<i>APOE</i> $\epsilon 4$ status (-/+)	-201, + 354 (0.64)	- 568, + 296 (0.34)	$< 0.0001$

ROS/MAP, The Religious Orders Study/the Rush Memory and Aging Project; ADNI, Alzheimer's Disease Neuroimaging Initiative; AD, Alzheimer's disease; Diff, the statistical difference between AD and non-AD; F, female; M, male; SD, standard deviation; Age at AD, age when AD developed for AD patients or age at last valid record for non-AD subjects; *APOE*  $\epsilon 4$  status (-/+), presence of the  $\epsilon 4$  allele

<sup>a</sup>Non-AD in ROS/MAP includes 713 individuals with missing AD status

<sup>b</sup>P values are calculated by Fisher's exact tests (for sex and *APOE*  $\epsilon 4$  status) or two-sample t tests (for age at death, age at AD)



avoid multicollinearity, only three dummy variables for four batches were used as covariates.

We employed three widely used models to test for significant genetic interactions in AD. The first method uses logistic regression models by including an additional interaction term [32]. The other two methods are faster in scanning for epistasis based on the inspection of  $3 \times 3$  joint genotype count tables. Boolean operation-based testing and screening (BOOST) allows the use of fast logic (bitwise) operations to obtain contingency tables [34]. Joint-effect tests maintain correct type 1 error rates under the null hypothesis [35]. For logistic regression, BOOST, and joint-effects, SNP pairs with fewer than 3 observations in any  $3 \times 3 \times 2$  contingency table cell (cases and controls were considered separately) were removed from the analysis, resulting in 392,241,651 valid tests being performed. Multiple-testing correction for statistical tests across the three methods was conducted using false discovery rate (FDR). Post hoc analysis was performed using a genotypic test [36] adjusting for sex, age, imputation batches, and the first three principal components to ensure that these potential confounders had not caused any observed association. Then, linkage disequilibrium (LD) between each pair of genetic interaction was measured by  $R^2$ . Genetic interactions with an  $LD > 0.2$  were removed from further analysis.

The associations of candidate interactions with AD pathologies and neuroimaging phenotypes were analyzed using the same genotypic test [36] adjusted for sex, age, imputation batches, and the first three principal components.

#### eQTL analysis and gene co-expression

Genetic interactions were further examined by eQTL and co-expression analysis. For eQTL analysis, data were obtained and analyzed using the Genotype-Tissue Expression (GTEx) web platform [37]. For co-expression analysis, SNPs were mapped to the nearest genes within a distance of 5 kb. RNA-Seq data were obtained from samples of the gray matter of the dorsolateral prefrontal cortex of 724 subjects from the ROS/MAP cohorts. These samples were quantified by using a Nanodrop spectrophotometer, and their quality was evaluated with an Agilent Bioanalyzer (Agilent Technologies, Inc., Santa Clara, CA, USA). A total of 582 RNA-Seq samples met the quality (Bioanalyzer RNA integrity (RIN) score  $> 5$ ) and quantity ( $5 \mu\text{g}$ ) thresholds. Then, the RNA-Seq data were processed via a parallelized automatic pipeline. The FPKMs were quantile normalized, and potential batch effects were removed by using the combat package in R. Pairwise correlations for gene co-expression were measured with the Pearson correlation coefficient.

#### Definition of epistasis risk scores and combined risk scores

PRSs have been used to inform the disease risk of a patient for the early prevention of the disease [17, 38]. Given a set of SNPs, PRSs are derived by multiplying the number of risk alleles for each SNP by the natural logarithm of their respective odds ratios (ORs) and summing these products for each subject [39]. For the  $j^{\text{th}}$  individual, the PRS is defined by:

$$PRS_j = \sum_i \frac{G_{ij} \times E_i}{N_j}$$

where  $E_i$  is the effect size for the  $i^{\text{th}}$  SNP,  $G_{ij}$  is the number of effective alleles observed for the  $i^{\text{th}}$  SNP of  $j^{\text{th}}$  individual, and  $N_j$  is the number of SNPs included in the PRS for the  $j^{\text{th}}$  individual.

For risk scores defined by epistasis, there is no readily available definition from previous publications. Inspired by an epistasis analysis framework called multifactor dimensionality reduction (MDR) [40, 41], we define the epistasis risk score (ERS) for the  $j^{\text{th}}$  individual as:

$$ERS_j = \sum_i \frac{\sum_{k=1}^9 C_{ik} E_{ik}}{N_j}$$

where  $C_{ik}$  equals to 1 if the  $j^{\text{th}}$  individual was assigned to the  $k^{\text{th}}$  cell of the  $3 \times 3$  genotype contingency table for the  $i^{\text{th}}$  interaction. Otherwise,  $C_{ik}$  equals to 0.  $E_{ik}$  is the effect size (natural logarithm of OR) of the  $k^{\text{th}}$  cell of  $3 \times 3$  genotype contingency table for the  $i^{\text{th}}$  interaction.  $N_j$  is the number of interactions included in the ERS for the  $j^{\text{th}}$  individual.

The combined risk score (CRS) of SNPs and SNP-SNP interactions for the  $j^{\text{th}}$  individual is defined as:

$$CRS_j = w \times PRS_j + (1 - w) \times ERS_j$$

where  $w$  is a weighting factor for PRS and ERS. To avoid using an arbitrary  $w$ , for individuals from ADNI, we selected  $w$  that maximized the AUC of CRSs in ROS/MAP. Likewise, we selected  $w$  that maximized the AUC of CRSs in ADNI for individuals from ROS/MAP.

#### Evaluation of AD risk using genetic interactions

We constructed several sets of risk scores to evaluate the contribution of epistasis to disease risk. For PRSs, *APOE* (rs7412 and rs429358) and 20 common SNPs identified by a previous GWAS [5] were included in the analysis. The effect size of each SNP was also obtained from the same study [5]. For ERSs, the effect size (natural logarithm of odds ratio) of each cell in the  $3 \times 3$  genotype contingency table for each genetic interaction was obtained on the discovery data ( $N = 10,389$ ). Based

on  $P$  value cutoff of  $1 \times 10^{-7}$ ,  $1 \times 10^{-6}$ , or  $1 \times 10^{-5}$ , three sets of ERSs were constructed. Moreover, the predictive power of each nominal significant genetic interaction ( $P < 1 \times 10^{-5}$ ) was evaluated via a permutation test. For each genetic interaction, an MDR model [41] was trained using the discovery data. Based on the MDR model, predictions of AD status were given for each genetic interaction on the two testing datasets, i.e., ROS/MAP and ADNI. The predictions for each interaction were permuted for 10,000 times in each testing dataset to generate a null distribution of random predictions. Based on the null distribution, we selected 77 interactions that showed non-random effects ( $P < 0.05$ ) in both ROS/MAP and ADNI.

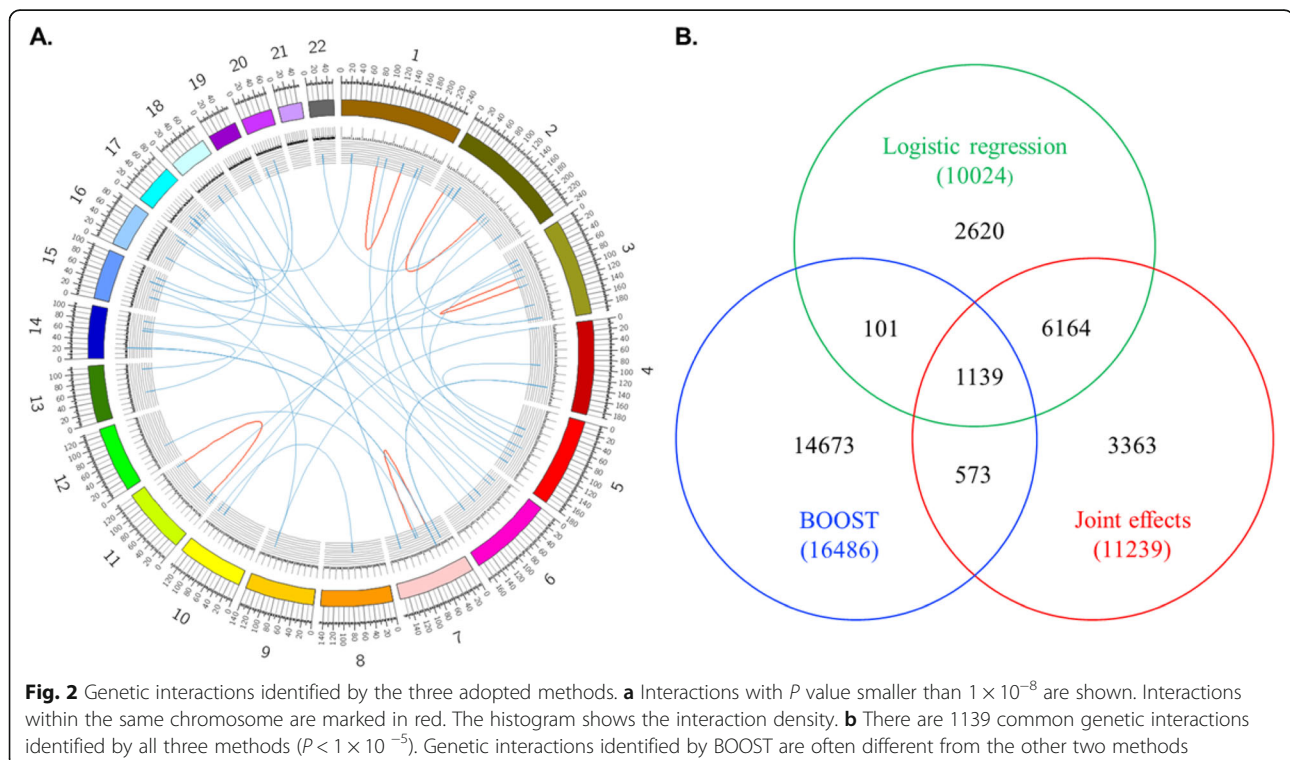
To evaluate whether the genetic risk scores were associated with age at onset of AD, we divided samples into 4 quantiles based on the corresponding genetic risk scores. The age at onset of AD in each quantile was analyzed with the Kaplan-Meier method, where patients were censored at the last record entry. The differences in the age at onset of AD in 4 quantiles were compared statistically using the log-rank test. Furthermore, the receiver operating characteristic (ROC) curves were generated by plotting the true-positive rate against the false-positive rate. Then, the area under the ROC curve (AUC) was calculated for each ROC curve to quantify the prediction accuracy of each type of genetic risk score.

## Results

### Genome-wide epistasis screening

First, we analyzed the main effect for each SNP using traditional genome-wide association analysis (GWAS). No genomic inflation was observed on the Q-Q plot (Figure S1). One genome-wide significant signal appeared nearby *APOE* on chromosome 19 (Figure S1). Epistasis screening for the clinical diagnosis of AD was carried out using three different methods. Under the BOOST model, a total of 16,486 SNP-SNP interactions (Table S1) were retained under a nominal  $P$  threshold of  $1 \times 10^{-5}$ . For logistic regression and joint-effect tests, 11,239 and 10,024 SNP-SNP interactions (Tables S2 and S3) were retained under a nominal  $P$  threshold of  $1 \times 10^{-5}$ , respectively. Overall, there were 28,633 SNP-SNP interactions identified by three methods (Fig. 2a). The SNP-SNP interactions obtained via different methods showed distinct results, with a higher overlap between logistic regression and joint effect tests (Fig. 2b).

Two significant SNP-SNP interactions were identified ( $P_{FDR} < 0.05$ ) by logistic regression. The first interaction was between rs6952399 (chr7: 45210711, *RAMP3*: intron variant) and rs6974494 (chr7: 83743961, *SEMA3A*: intron variant). The other was between rs17856580 (chr16: 27246617, *NSMCE1*: missense variant) and rs1048159 (chr17: 54912339, *DGKE*: synonymous variant, *C17orf67*: upstream variant). Under a less conservative  $P_{FDR}$  of 0.1, the interaction between rs2164808 (chr2: 25377176, *EFR3B*: stop gained, *RP11-509E16.1*: upstream variant)



and rs354709 (chr2:143886953, *ARHGAP15*: intron variant) was the only genetic interaction identified by BOOST. Five genetic interactions ( $P_{FDR} < 0.1$ ) were identified by joint effect analysis, including rs6952399-rs6974494 which was also identified by logistic regression. No main effects were observed for these genetic interactions with an  $P_{FDR} < 0.1$ , except a nominal significant signal ( $P = 0.017$ ) for rs2301600 in rs4574537-rs2301600. Interactions ( $P_{FDR} < 0.1$ ) identified by three methods were displayed in Table 3.

### Transcription analysis of candidate interactions

The expression levels of genes that interact with each other are likely to be positively or negatively correlated [42]. Combining the interaction pattern with co-expression and eQTL analysis, we can gain biological insight beyond the statistical significance. Visualization of the genetic interaction showed that rs6952399<sup>G</sup>-rs6974494<sup>TT</sup> carriers displayed a higher risk of developing AD (Fig. 3). Based on eQTL analysis, rs6952399<sup>G</sup> carrier showed a higher expression of *RAMP3* ( $P = 2.4 \times 10^{-6}$ ), and rs6974494<sup>T</sup> carrier showed a lower expression of *SEMA3A* ( $P = 1.6 \times 10^{-5}$ ). Therefore, it is likely that the upregulation of *RAMP3* expression combined with the downregulation of *SEMA3A* confers a higher risk of AD (Fig. 3). This assumption is further supported by the fact that *RAMP3* and *SEMA3A* showed negative co-expression in the brain ( $R = -0.29$ ). Moreover, *RAMP3* and *SEMA3A* demonstrated higher negative correlation in AD patients ( $R = -0.33$ ) compared to cognitive normal controls ( $R = -0.24$ ).

The seven SNP-SNP interactions identified by the three models correspond to nine gene-gene interactions, five of which showed significant co-expression. Besides the co-expression between *RAMP3* and *SEMA3A*, a negative correlation has been observed between *NSMC*

*E1* and *DGKE/C17orf67* and between *ARHGAP15* and *EFR3B/RP11-509E16.1* as well (Table S4). Statistical epistasis often lacks biological interpretation. Here, we provide the visualization of genetic interactions (Fig. 3, Figure S2) together with the results of co-expression and eQTL analysis (Table S4) to facilitate biological interpretation of statistical significance.

### Candidate interactions and AD neuropathology

To investigate the biological mechanism of the identified genetic interactions, we tested the associations of these interactions ( $P_{FDR} < 0.1$ ) with intermediate AD phenotypes including A $\beta$  and tau protein levels, brain atrophy, and white matter injury. None of the candidate genetic interactions displayed a significant association with the entorhinal cortex or hippocampal volume. One interaction, between rs8580 (chr7: 44620836, *TMED4*: synonymous variant) and rs8004063 (chr14: 23732479, *C14orf164*: intron variant), demonstrated significant association with neurite plaques in the entorhinal cortex ( $P = 0.019$ ). Due to linkage disequilibrium between rs217362 and rs8580, the rs217362-rs8004063 interaction also manifested a significant association with neurite plaques in the entorhinal cortex ( $P = 0.023$ ). *TMED4* belongs to p24 family proteins, which are mainly involved in vesicular protein trafficking and are likely to promote neuritic plaque formation in AD [43]. *C14orf164* is an important paralog of *RNF212* which can encode an E3 enzyme in the ubiquitin proteasome system whose dysfunction could lead to A $\beta$  accumulation [44, 45]. Their molecular function supports that the interaction may act through the trafficking and processing of APP in AD pathogenesis.

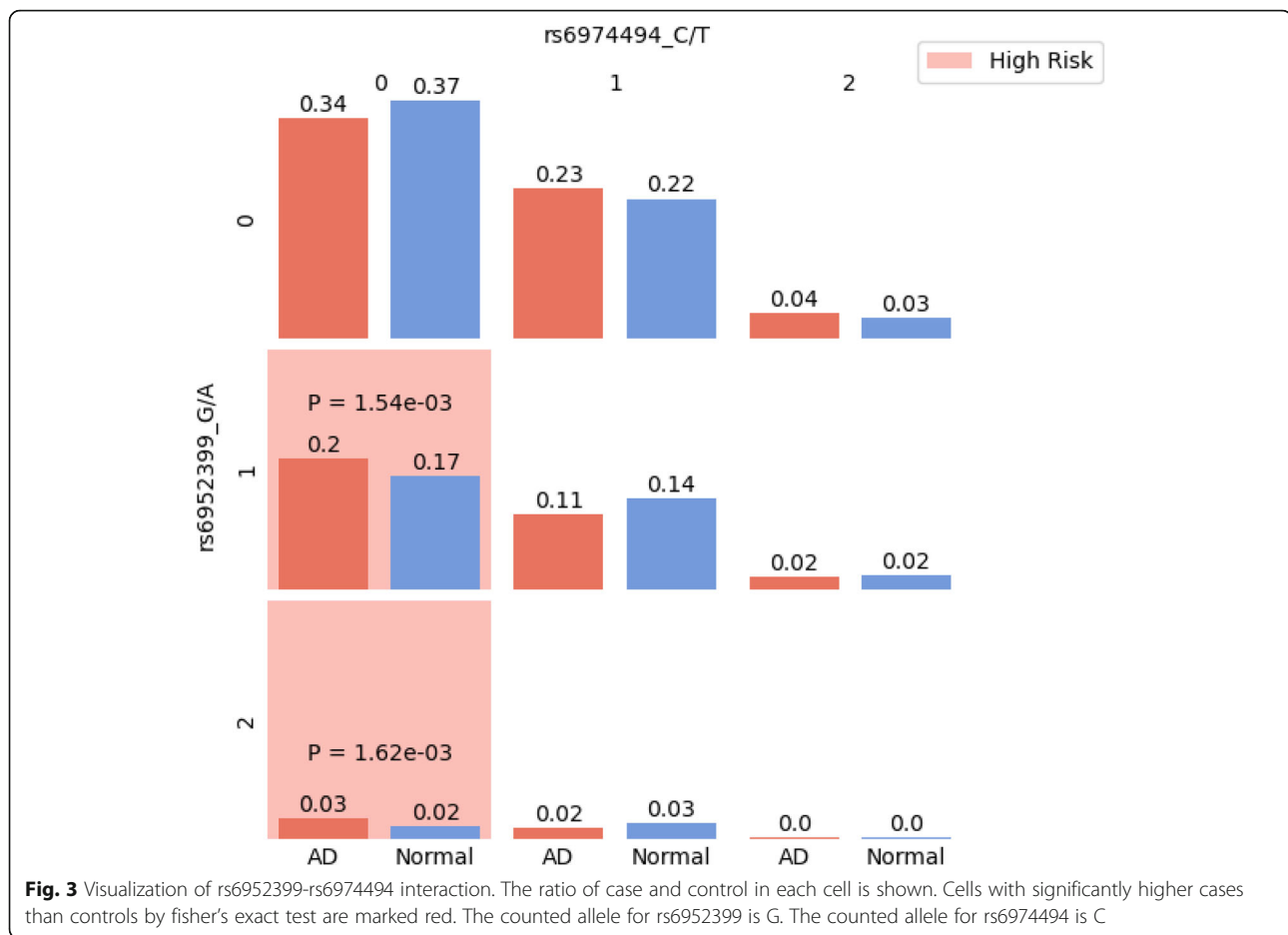
White matter (WM) fractional anisotropy (FA) is thought to be related to WM integrity, and a decline in FA is often used as an index of decreasing WM health.

**Table 3** Significant genetic interactions with a false discovery rate (FDR) < 0.1

SNP1	Chr_1	Position_1 <sup>a</sup>	SNP2	Chr_2	Position_2 <sup>a</sup>	P	FDR	Gene_1	Gene_2
<b>Genetic interaction identified by BOOST</b>									
rs354709	2	143886953	rs2164808	2	25377176	1.47E-10	0.058	<i>ARHGAP15</i>	<i>EFR3B; RP11-509E16.1</i>
<b>Genetic interactions identified by logistic regression</b>									
rs6952399	7	45210711	rs6974494	7	83743961	1.88E-10	0.048	<i>RAMP3</i>	<i>SEMA3A</i>
rs17856580	16	27246617	rs1048159	17	54912339	2.43E-10	0.048	<i>NSMCE1</i>	<i>DGKE; C17orf67</i>
<b>Genetic interactions identified by joint effect</b>									
rs6952399	7	45210711	rs6974494	7	83743961	5.08E-10	0.076	<i>RAMP3</i>	<i>SEMA3A</i>
rs217362	7	44618599	rs8004063	14	23732479	7.61E-10	0.076	<i>DDX56; TMED4</i>	<i>C14orf164; RNU6-1046P</i>
rs4574537	5	137419728	rs2301600	19	35786868	7.70E-10	0.076	<i>WNT8A</i>	<i>MAG</i>
rs2075302	2	163076146	rs374479	5	80998915	8.03E-10	0.076	<i>FAP</i>	<i>SSBP2</i>
rs8580	7	44620836	rs8004063	14	23732479	9.70E-10	0.076	<i>TMED4</i>	<i>C14orf164; RNU6-1046P</i>

Chr\_1, chromosome of SNP1; Gene\_1, genes within 5 kb of SNP1; Chr\_2, chromosome of SNP2; Gene\_2, genes within 5 kb of SNP2;

<sup>a</sup>SNP positions were in build 37, assembly hg19



The interaction between rs2164808 (chr2: 25377176, *EFR3B*: stop gained, *RP11-509E16.1*: upstream variant) and rs354709 (chr2: 143886953, *ARHGAP15*: intron variant) showed significant associations (left  $P = 0.046$ , right  $P = 0.023$ ) with FA estimates in the splenium of the corpus callosum. *EFR3B* and phosphatidylinositol 4-kinase alpha (*PI4KA*) forms a protein complex that plays an important role in the myelination process via actin dynamics [46]. Interestingly, *ARHGAP15*, which is a Rac-specific negative regulator, is also heavily involved in actin cytoskeleton dynamics [47], suggesting that its interaction with *EFR3B* affects WM health in AD patients through regulating myelination in axons.

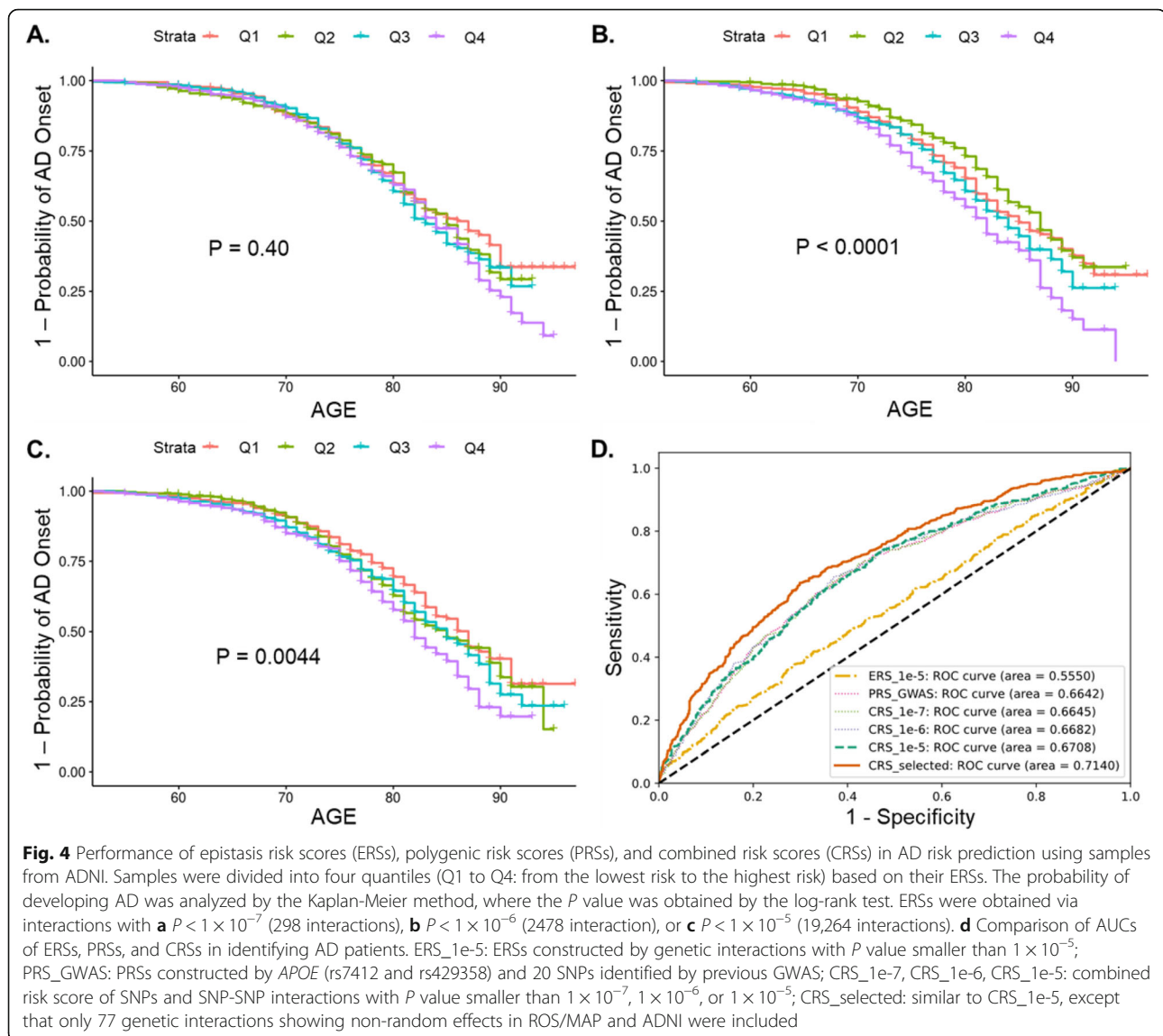
#### Epistasis risk scores in AD

The polygenic basis of LOAD can be harnessed to identify individuals at risk for cognitive decline. Previously, PRSs were inferred from the cumulated effects of each disease-associated SNP. We investigated whether ERSs that were inferred from the summed effects of each disease-associated SNP-SNP interactions could serve as an indicator of disease risk. After removing redundant genetic interactions due to LD, 19,264 of 28,633 genetic

interactions with  $P < 1 \times 10^{-5}$  were kept for ERS analysis. ERSs were constructed based on three different  $P$  value thresholds:  $P < 1 \times 10^{-7}$  (298 interactions),  $P < 1 \times 10^{-6}$  (2478 interactions), and  $P < 1 \times 10^{-5}$  (19,264 interactions).

We evaluated if individuals with higher ERSs had a higher risk of AD, therefore, had onset of AD at an earlier age. It was shown that ERSs constructed from genetic interactions with  $P < 1 \times 10^{-7}$  could not identify high-risk individuals from low-risk individuals in either ADNI or ROS/MAP (Fig. 4a; Figure S3). This may arise from the fact that SNPs or SNP-SNP interactions identified from association analysis often had a very small effect size. Adding more genetic interactions (e.g., 2478 interactions or 19,264 interactions) gradually increased the power of ERSs to stratify high risk individuals from low risk individuals (log-rank test  $P < 0.0001$ , Fig. 4b; log-rank test  $P = 0.0044$ , Fig. 4c; Figure S3). Thus, we demonstrated that ERS could serve as an indicator of the genetic risk of AD (Fig. 4; Figure S3). The same conclusion still held true when samples from ROS/MAP were used (Figure S3). Furthermore, we evaluated if the predictive power of ERSs was due to SNPs that have a main





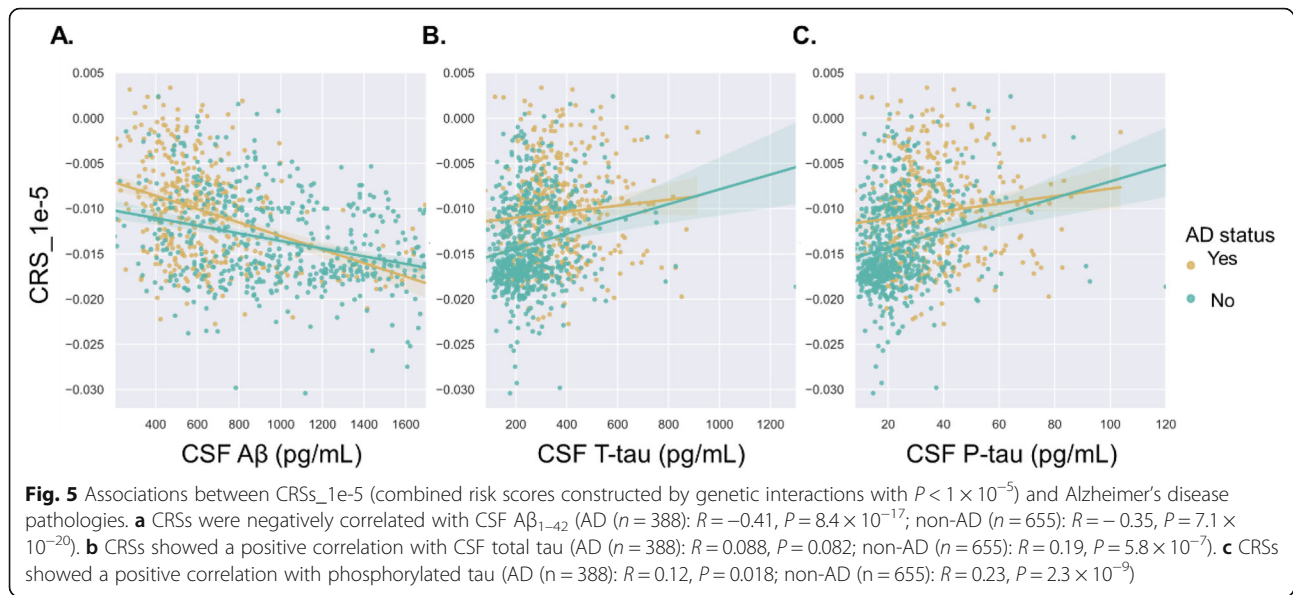
effect in AD. We found that ERSs still demonstrated the power to stratify high-risk individuals from low-risk individuals, after removing interactions that contained any main effect SNP ( $P < 0.05$ ) (Figure S4).

#### Combined risk scores of SNPs and SNP-SNP interactions

We evaluated if combined risk scores (CRS) of SNPs (defining PRS) and SNP-SNP interactions (defining ERS) could be a better indicator of AD risk. PRSs and ERSs showed an AUC of 0.66 and 0.56, respectively (Fig. 4d). When ERSs consisting of genetic interactions with  $P < 1 \times 10^{-7}$ ,  $1 \times 10^{-6}$ , or  $1 \times 10^{-5}$  were combined with PRSs for the construction of CRSs, the AUC of CRSs showed a non-significant but steady increase up to 0.67 (Fig. 4d). The similar non-significant but steady increase of AUC for CRSs obtained using different  $P$  value thresholds

were also detected using individuals from ROS/MAP (Figure S3).

We evaluated the correlation between CRSs derived from genetic interactions with  $P < 1 \times 10^{-5}$  and CSF markers of AD. It was found that CRSs displayed a strong negative correlation with CSF  $\beta$ -amyloid (1–42) ( $R = -0.40$ ,  $P = 1.8 \times 10^{-35}$ ) and a strong positive correlation with CSF total tau ( $R = 0.24$ ,  $P = 2.2 \times 10^{-15}$ ) and CSF phosphorylated tau ( $R = 0.27$ ,  $P = 2.1 \times 10^{-19}$ ) (Fig. 5). Interestingly, CRSs showed much stronger correlation with CSF total tau (AD:  $R = 0.088$ ,  $P = 0.082$ ; non-AD:  $R = 0.19$ ,  $P = 5.8 \times 10^{-7}$ ) and CSF phosphorylated tau (AD:  $R = 0.12$ ,  $P = 0.018$ ; non-AD:  $R = 0.23$ ,  $P = 2.3 \times 10^{-9}$ ) in cognitive normal controls than in AD patients (Fig. 5b, c). However, a higher correlation between CRSs and CSF  $\beta$ -amyloid (1–42) (AD:  $R = -0.41$ ,  $P = 8.4 \times 10^{-17}$ ; non-AD:  $R = -0.35$ ,  $P = 7.1 \times 10^{-20}$ ) was observed



in AD patients (Fig. 5a). This also held true when PRSs were used in the correlation analysis (Figure S5). These results strongly suggest that tau may act in the earlier stage of AD as high-risk individuals showed faster accumulation of CSF tau when they still displayed normal cognitive status.

It should be noted that CRSs did not display a higher correlation with CSF markers compared with PRSs (Fig. 5; Figure S5). A previous study has discovered that APOE contributed mostly to amyloid accumulation and other SNPs only affected the risk of further conversion to AD [48]. The present analysis also reveals that genetic interactions affected the risk of AD rather than pathological markers.

Furthermore, we selected interactions that demonstrated the power to predict the risk of AD by itself alone (Methods). From 19,264 interactions with  $P < 1 \times 10^{-5}$ , 77 interactions that displayed predictive effects in both ROS/MAP and ADNI were selected. Combined risk scores of SNPs and these 77 interactions performed much better in predicting the clinical status of AD (ADNI: AUC = 0.71, Fig. 4d; ROSMAP: AUC = 0.69, Figure S3). However, because the testing datasets (ADNI and ROS/MAP) were used for selection, the prediction accuracy of this subset of interactions should be interpreted with caution. Their effects on risk prediction need to be evaluated when a new dataset is available. Overall, they represented a subset of SNP-SNP interactions, each of which alone demonstrated predictive capacity across our discovery and testing data. The full list of the selected interactions is shown in Table S5.

## Discussion

To help explain the missing heritability in AD, we performed a genome-wide interaction analysis of AD. There

were seven candidate genetic interactions ( $P_{FDR} < 0.1$ ) identified using the three most popularly adopted methods. Previous reports supported possible functional convergence between pairs of genes identified by our analysis, such as *RAMP3-SEMA3A* and *NSMCE1-DGKE*. It was reported that *RAMP3*, a component of amylin receptor-3, could induce cell death via neurotoxic actions of  $A\beta$  [49]. Semaphoring 3A (*Sema3A*) could bind to nonamyloidogenic sAPP $\alpha$  which would prevent the collapse of axonal growth cones induced by *Sema3A* [50]. Consequently, biological interaction between *RAMP3* and *Sema3A* is likely to be involved in the neurodegeneration process of AD. For *NSMCE1* and *DGKE*, there may exist direct physical interaction between them, as two independent studies have uncovered the exact same ubiquitination site at lysine 357 in human *DGKE* [51, 52]. *NSMCE1* is a RING-type zinc finger-containing E3 ubiquitin ligase that assembles with melanoma antigen protein to catalyze the direct transfer of ubiquitin from E2 ubiquitin-conjugating enzyme to a specific substrate. *DGKE* is a membrane-bound diacylglycerol kinase that converts diacylglycerol into phosphatidic acid.

Moreover, we visualized the  $3 \times 3$  contingency table of each interaction (Fig. 3; Figure S2). We attempted to combine the observed interaction pattern with the gene expression pattern (i.e., co-expression and eQTL) to infer the mechanism of action of each interaction. In this way, we found that the higher expression of *RAMP3* combined with the lower expression of *SEMA3A* conferred a higher risk of AD. Then, we related candidate interactions with intermediate phenotypes in AD such as  $A\beta$  and tau levels, brain atrophy, and FA estimates to help understand the biological consequences of the statistical significance. The association between *TMED4*-

*C14orf164* and neurite plaques in the entorhinal cortex indicates that ubiquitination may play an important role in A $\beta$  accumulation, as *C14orf164* is an important paralog of *RNF212* which can encode an E3 enzyme in the ubiquitin proteasome.

Epistasis has never been used in the construction of genetic risk scores. Here, we demonstrated that ERSs were able to discriminate high-risk individuals that were more likely to develop AD (Fig. 4; Figure S3). Then, combined risk scores of SNP and SNP-SNP interactions showed slightly but steadily increased AUC in predicting the clinical status of AD (Fig. 4d). Additionally, we selected a subset of 77 genetic interactions that showed non-random effects in both ROS/MAP and ADNI. It was shown that combined risk scores including the 77 interactions performed better in predicting the clinical status of AD than using all the genetic interactions with  $P < 1 \times 10^{-5}$  (Fig. 4d; Figure S3). This indicated the possibility of combining PRSs and ERSs as potential biomarkers of AD. However, further evaluation of the selected interactions on new datasets is needed. Altogether, we demonstrated that ERS is a promising complement to traditional PRS in practical application.

### Limitations

To reduce the search space, we only analyzed SNPs with a CADD score  $\geq 15$  that were more likely to be causative. However, it is still possible for two non-deleterious SNPs to be disease-causing variants when there is a genetic interaction between them. In that case, faster tests are needed to test the interactions between millions of SNPs. Moreover, tests for interactions are complicated by the fact that samples are stratified by the  $3 \times 3$  genotype contingency table. Therefore, cells with very small sample sizes are likely to induce false positives in the test results. We try to avoid this issue by removing pairs that show few samples in any cells of the  $3 \times 3$  genotype contingency table. Thus, we are likely to have removed some rare allele pairs that might interact with each other. Overall, the predictive power of our model was based on a selected subset of deleterious common variants; further improvement may be expected when non-deleterious or rare variants could be incorporated.

Moreover, we simply combined PRSs and ERSs by a weighting factor. There may be a more complicated relationship between the additive effects of single SNPs and genetic interactions. The development of a better integrated model that can account for both main effects and epistasis would further increase the prediction accuracy of the genetic risk score.

### Conclusions

In conclusion, through a genome-wide epistasis analysis, we identified a number of genetic interactions that are

often co-expressed and can partly explain the “missing heritability” in AD. Subsequent analysis revealed possible links between these genetic interactions and pathological endophenotypes. Furthermore, it was demonstrated that ERSs can identify high-risk individuals showing earlier onset of AD. Combined risk scores of SNPs and SNP-SNP interactions showed slightly but steadily increased AUC in predicting the clinical status of AD.

### Abbreviations

A $\beta$ :  $\beta$ -amyloid; AD: Alzheimer disease; ADC: National Institute on Aging Alzheimer's Disease Center; ADNI: Alzheimer's Disease Neuroimaging Initiative; AUC: Area under the receiver operating characteristic curve; BOOST: Boolean operation-based testing and screening; CADD: Combined annotation-dependent depletion; CRS: Combined risk score; CSF: Cerebrospinal fluid; eQTL: Expression quantitative trait loci; ERS: Epistasis risk score; GenADA: Genetic Alzheimer's Disease Associations; GWAS: Genome-wide association studies; LD: Linkage disequilibrium; LOAD: Late-onset Alzheimer disease; NIA-LOAD: National Institute on Aging-Late-Onset-Alzheimer's-Disease Study; PRS: Polygenic risk score; ROS/MAP: The Religious Orders Study and the Rush Memory and Aging Project; SNP: Single nucleotide polymorphisms

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13195-021-00794-8>.

**Additional file 1: Figure S1.** GWAS analysis of selected SNPs (N = 10389, SNPs = 36860). **Figure S2.** Visualization of SNP-SNP interactions. **Figure S3.** Performance of epistasis risk scores (ERSs), polygenic risk scores (PRSs) and combined risk scores (CRSs) in AD risk prediction using samples from ROS/MAP. **Figure S4.** Epistasis risk scores (ERSs) analysis after removing genetic interactions that showed main effects ( $P < 0.05$ ) using samples from ADNI and ROS/MAP. **Figure S5.** Associations between PRSs (polygenic risk scores constructed by APOE, i.e., rs7412 and rs429358 and 20 SNPs identified by previous GWAS) and Alzheimer's disease pathologies.

**Additional file 2: Table S1.** Nominal significant SNP-SNP interactions (BOOST). **Table S2.** Nominal significant SNP-SNP interactions (Logistic regression). **Table S3.** Nominal significant SNP-SNP interactions (Joint effect). **Table S4.** Co-expression and eQTL analysis of significant genetic interactions (FDR < 0.1). **Table S5.** Selected genetic interactions that showed non-random prediction in both ADNI and ROSMAP.

### Acknowledgements

We thank the dbGaP, ROS/MAP, ADNI for providing the study materials, clinical information, and technical support.

### Financial disclosures

Hui Wang reports no disclosures.  
David A. Bennett reports no disclosures.  
Philip L De Jager reports no disclosures.  
Qing-Ye Zhang reports no disclosures.  
Hong-Yu Zhang reports no disclosures.

### Statistical analysis

Conducted by Hui Wang, Ph.D., Huazhong Agricultural University

### Authors' contributions

HW developed the method and performed the analyses. HW and HYZ wrote the manuscript. HYZ designed and conceptualized the study. DAB and QYZ contributed to the data acquisition. DAB, PLDJ, and HYZ verified the analytical methods and revised the manuscript. The authors read and approved the final manuscript.

## Funding

This work was supported by the National Natural Science Foundation of China (Grant 31670779) and the Fundamental Research Funds for the Central Universities (Grant 2662017PY115).

Work from Rush was supported in part by grants P30AG10161, R01AG15819, R01AG17917, U01AG61356, R01AG30146, the Illinois Department of Public Health, and the Translational Genomics Research Institute (Kronos Science). We are indebted to the participants in the Religious and the Orders Study and the Rush Memory and Aging Project.

Data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database were 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). The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report.

## Availability of data and materials

Genotypes and phenotypes of samples from NIA-LOAD, GenADA, and ADCS can be accessed from dbGaP database (<https://dbgap.ncbi.nlm.nih.gov>) under the accession number of phs000168.v2.p2, phs000219.v1.p1, and phs000372.v1.p1, respectively. Genotypes, gene expression profiles, and post-mortem pathological measurements of AD patients from ROS/MAP can be obtained through RADC (<https://www.radc.rush.edu>). Genotypes, CSF pathological measurements, and neuroimages of samples from ADNI can be applied from <http://adni.loni.usc.edu>.

## Declarations

### Ethics approval and consent to participate

The usage of three datasets from dbGaP was approved by the institutional review board at the Huazhong Agricultural University. Data from the ROS/MAP were obtained under the data use agreement at Rush Alzheimer's Disease Center (RADC). The usage of ADNI samples follows the data use agreement at ADNI. All participants gave written informed consent.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Huazhong Agricultural University, College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, No.1 Shizishan Street, Hongshan District, Wuhan 430070, Hubei, China. <sup>2</sup>Rush University Medical Center, Rush Alzheimer's Disease Center, Chicago, IL, USA. <sup>3</sup>Rush University Medical Center, Department of Neurological Sciences, Chicago, IL, USA. <sup>4</sup>Columbia University Medical Center, Center for Translational and Computational Neuroimmunology, New York, NY, USA. <sup>5</sup>Broad Institute, Cell Circuits Program, Cambridge, MA, USA.

Received: 10 July 2020 Accepted: 22 February 2021

Published online: 04 March 2021

## References

- Qiu C, Kivipelto M, von Strauss E. Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. *Dialogues Clin Neurosci*. 2009;11:111–28.
- Hardy JA, Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. *Science*. 1992;256:184–6.
- 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;51:404–13.
- Lambert J-C, 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;45:1452–8.
- 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 A $\beta$ , tau, immunity and lipid processing. *Nat Genet*. 2019;51:414–30.
- Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Consortium ADG. Alzheimer's disease: analyzing the missing heritability. *PLoS One*. 2013;8:e79771.
- Lee SH, Harold D, Niyholt DR, ANZGene Consortium, International Endogene Consortium, Genetic and Environmental Risk for Alzheimer's disease Consortium, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet*. 2013;22:832–41.
- 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;63:168–74.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747.
- Vardarajan BN, Zhang Y, Lee JH, Cheng R, Bohm C, Ghani M, et al. Coding mutations in SORL1 and Alzheimer disease. *Ann Neurol*. 2015;77:215–27.
- Steinberg S, Stefansson H, Jonsson T, Johannsdottir H, Ingason A, Helgason H, et al. Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. *Nat Genet*. 2015;47:445–7.
- 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;49:1373–84.
- Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. *Front Genet*. 2015;6:285.
- Johnstone IM, Titterton DM. Statistical challenges of high-dimensional data. *Philos Trans A Math Phys Eng Sci*. 2009;367:4237–53.
- Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 2002;11:2463–8.
- Gusareva ES, Carrasquillo MM, Bellenguez C, Cuyvers E, Colon S, Graff-Radford NR, et al. Genome-wide association interaction analysis for Alzheimer's disease. *Neurobiol Aging*. 2014;35:2436–43.
- Sleegers K, Bettens K, De Roeck A, Van Cauwenberghe C, Cuyvers E, Verheijen J, et al. A 22-single nucleotide polymorphism Alzheimer's disease risk score correlates with family history, onset age, and cerebrospinal fluid A $\beta$ 42. *Alzheimers Dement*. 2015;11:1452–60.
- Felsky D, Patrick E, Schneider JA, Mostafavi S, Gaiteri C, Patsopoulos N, et al. Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. *Mol Neurodegener*. 2018;13:38.
- Mormino EC, Sperling RA, Holmes AJ, Buckner RL, De Jager PL, Smoller JW, et al. Polygenic risk of Alzheimer disease is associated with early-and late-life processes. *Neurology*. 2016;87:481–8.
- Lee JH, Cheng R, Graff-Radford N, Foroud T, Mayeux R. Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. *Arch Neurol*. 2008;65:1518–26.
- Filippini N, Rao A, Wetten S, Gibson RA, Borrie M, Guzman D, et al. Anatomically-distinct genetic associations of APOE  $\epsilon$ 4 allele load with regional cortical atrophy in Alzheimer's disease. *Neuroimage*. 2009;44:724–8.
- Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet*. 2011;43:436.
- Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64:S161–89.
- Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, et al. Alzheimer's disease neuroimaging initiative (ADNI): clinical characterization. *Neurology*. 2010;74:201–9.
- Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshev YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48:1443.
- Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284.
- Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, et al. The UCSC genome browser database: 2018 update. *Nucleic Acids Res*. 2017;46:D762–9.
- Van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Abdellaoui A, Hofman A, et al. Population-specific genotype imputations using minimac or IMPUTE2. *Nat Protoc*. 2015;10:1285.
- Consortium GP, Auton A, Brooks LD. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
- Jin Y, Schaffer AA, Feolo M, Holmes JB, Kattman BL. GRAF-pop: a fast distance-based method to infer subject ancestry from multiple genotype



- datasets without principal components analysis. *G3 Bethesda Md.* 2019;9:2447–61.
31. Rentsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* 2018;47:D886–94.
  32. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
  33. Jiang L, Zheng Z, Qi T, Kemper KE, Wray NR, Visscher PM, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat Genet.* 2019;51:1749–55.
  34. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, et al. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet.* 2010;87:325–40.
  35. Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. *PLoS Genet.* 2012;8:e1002625.
  36. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics.* 2009;25:3275–81.
  37. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The genotype-tissue expression (GTEx) project. *Nat Genet.* 2013;45:580.
  38. Tan CH, Bonham LW, Fan CC, Mormino EC, Sugrue LP, Broce IJ, et al. Polygenic hazard score, amyloid deposition and Alzheimer's neurodegeneration. *Brain.* 2019;142:460–70.
  39. Euesden J, Lewis CM, O'reilly PF. PRSice: polygenic risk score software. *Bioinformatics.* 2014;31:1466–8.
  40. Dai H, Charnigo RJ, Becker ML, Leeder JS, Motsinger-Reif AA. Risk score modeling of multiple gene to gene interactions using aggregated-multifactor dimensionality reduction. *BioData Min.* 2013;6:1.
  41. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet.* 2001;69:138–47.
  42. Wang X, Fu AQ, McNerney ME, White KP. Widespread genetic epistasis among cancer genes. *Nat Commun.* 2014;5:4828.
  43. Qiu K, Zhang X, Wang S, Li C, Wang X, Li X, et al. TMP21 in Alzheimer's disease: molecular mechanisms and a potential target. *Front Cell Neurosci.* 2019;13:328.
  44. Qiao H, Prasada Rao HBD, Yang Y, Fong JH, Cloutier JM, Deacon DC, et al. Antagonistic roles of ubiquitin ligase HEI10 and SUMO ligase RNF212 regulate meiotic recombination. *Nat Genet.* 2014;46:194–9.
  45. Oddo S. The ubiquitin-proteasome system in Alzheimer's disease. *J Cell Mol Med.* 2008;12:363–73.
  46. Alvarez-Prats A, Bjelobaba I, Aldworth Z, Baba T, Abebe D, Kim YJ, et al. Schwann-cell-specific deletion of phosphatidylinositol 4-kinase alpha causes aberrant myelination. *Cell Rep.* 2018;23:2881–90.
  47. Zamboni V, Armentano M, Berto G, Ciralo E, Ghigo A, Garzotto D, et al. Hyperactivity of Rac1-GTPase pathway impairs neuritogenesis of cortical neurons by altering actin dynamics. *Sci Rep.* 2018;8:7254.
  48. Leonenko G, Shoai M, Bellou E, Sims R, Williams J, Hardy J, et al. Genetic risk for Alzheimer disease is distinct from genetic risk for amyloid deposition. *Ann Neurol.* 2019;86:427–35.
  49. Fu W, Ruangkittisakul A, MacTavish D, Shi JY, Ballanyi K, Jhamandas JH. Amyloid  $\beta$  (A $\beta$ ) peptide directly activates amylin-3 receptor subtype by triggering multiple intracellular signaling pathways. *J Biol Chem.* 2012;287:18820–30.
  50. Magdesian MH, Gralle M, Guerreiro LH, Beltrão PJJ, Carvalho MMVF, Santos LE d S, et al. Secreted human amyloid precursor protein binds semaphorin 3a and prevents semaphorin-induced growth cone collapse. *PLoS One.* 2011;6:e22857.
  51. Mertins P, Qiao JW, Patel J, Udeshi ND, Clauser KR, Mani DR, et al. Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods.* 2013;10:634–7.
  52. Wagner SA, Beli P, Weinert BT, Schölz C, Kelstrup CD, Young C, et al. Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. *Mol Cell Proteomics.* 2012;11:1578–85.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://www.biomedcentral.com/submissions)

