Hong et al. *Translational Psychiatry* (2020)10:403
https://doi.org/10.1038/s41398-020-01074-z

Translational Psychiatry

ARTICLE Open Access

# Genome-wide association study of Alzheimer's disease CSF biomarkers in the EMIF-AD Multimodal Biomarker Discovery dataset

Shengjun Hong[1], Dmitry Prokopenko[2], Valerija Dobricic[1], Fabian Kilpert[1], Isabelle Bos[3,4], Stephanie J. B. Vos[3], Betty M. Tijms[5], Ulf Andreasson[6,7], Kaj Blennow[6,7], Rik Vandenberghe[8,9], Isabelle Cleynen[10], Silvy Gabel[8], Jolien Schaeverbeke[8], Philip Scheltens[5], Charlotte E. Teunissen[11], Ellis Niemantsverdriet[12], Sebastiaan Engelborghs[12,13], Giovanni Frisoni[14,15], Olivier Blin[16], Jill C. Richardson[17], Regis Bordet[18], José Luis Molinuevo[19], Lorena Rami[19], Alzheimer's Disease Neuroimaging Initiative (ADNI), Petronella Kettunen[6,20], Anders Wallin[6], Alberto Lleó[21], Isabel Sala[21], Julius Popp[22,23], Gwendoline Peyratout[23], Pablo Martinez-Lage[24], Mikel Tainta[24], Richard J. B. Dobson[25,26,27,28,29], Cristina Legido-Quigley[30,31], Kristel Sleegers[32,33], Christine Van Broeckhoven[32,33], Mara ten Kate[34,35], Frederik Barkhof[36], Henrik Zetterberg[6,7,37,38], Simon Lovestone[39], Johannes Streffer[40,41], Michael Wittig[42], Andre Franke[42], Rudolph E. Tanzi[2], Pieter Jelle Visser[5] and Lars Bertram[1,43]

## Abstract

Alzheimer's disease (AD) is the most prevalent neurodegenerative disorder and the most common form of dementia in the elderly. Susceptibility to AD is considerably determined by genetic factors which hitherto were primarily identified using case–control designs. Elucidating the genetic architecture of additional AD-related phenotypic traits, ideally those linked to the underlying disease process, holds great promise in gaining deeper insights into the genetic basis of AD and in developing better clinical prediction models. To this end, we generated genome-wide single-nucleotide polymorphism (SNP) genotyping data in 931 participants of the European Medical Information Framework Alzheimer's Disease Multimodal Biomarker Discovery (EMIF-AD MBD) sample to search for novel genetic determinants of AD biomarker variability. Specifically, we performed genome-wide association study (GWAS) analyses on 16 traits, including 14 measures derived from quantifications of five separate amyloid-beta (Aβ) and tau-protein species in the cerebrospinal fluid (CSF). In addition to confirming the well-established effects of apolipoprotein E (*APOE*) on diagnostic outcome and phenotypes related to Aβ42, we detected novel potential signals in the zinc finger homeobox 3 (*ZFHX3*) for CSF-Aβ38 and CSF-Aβ40 levels, and confirmed the previously described sex-specific association between SNPs in geminin coiled-coil domain containing (*GMNC*) and CSF-tau. Utilizing the results from independent case–control AD GWAS to construct polygenic risk scores (PRS) revealed that AD risk variants only explain a small fraction of CSF biomarker variability. In conclusion, our study represents a detailed first account of GWAS analyses on CSF-Aβ and -tau-related traits in the EMIF-AD MBD dataset. In subsequent work, we will utilize the genomics data generated here in GWAS of other AD-relevant clinical outcomes ascertained in this unique dataset.

Correspondence: Lars Bertram (lars.bertram@uni-luebeck.de)
[1]Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), Institutes of Neurogenetics and Cardiogenetics, University of Lübeck, Lübeck, Germany
[2]Genetics and Aging Unit and McCance Center for Brain Health, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
Full list of author information is available at the end of the article
Alzheimer's Disease Neuroimaging Initiative (ADNI): Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

## Introduction

Alzheimer's disease (AD) is a progressive and devastating neurodegenerative disorder, which leads to cognitive decline, loss of autonomy, dementia, and eventually death. Neuropathologically, AD is characterized by the accumulation of extracellular amyloid β (Aβ) peptide deposits ("plaques") and intracellular hyperphosphorylated tau protein aggregates ("tangles") in the

brain[1,2]. Using genetic linkage analysis followed by positional cloning led to the discovery of rare mutations in three genes encoding the amyloid-beta precursor protein (*APP*) and presenilins 1 and 2 (*PSEN1*, *PSEN2*) that cause fully penetrant monogenic forms of AD[3]. However, the vast majority of patients likely suffer from a polygenic ("sporadic") form of AD, which is driven by numerous genomic variants[4], the identification of which are the main aim of genome-wide association studies (GWAS).

The most strongly and most consistently associated AD risk gene (even prior to GWAS[5]) is *APOE*, which encodes apolipoprotein E, a cholesterol transport protein that has been implicated in numerous amyloid-specific pathways, including amyloid trafficking, as well as plaque clearance[2,6]. In addition to *APOE*, nearly three dozen independent loci have now been reported to be associated with disease risk by GWAS[7–9]. Pathophysiologically, the risk genes identified to date appear to predominantly act through modulations of the immune system response, endocytotic mechanisms, cholesterol homeostasis, and APP catabolic processes[7,8].

Despite these general advances in the field of AD genetics, many key questions still remain to be answered. First, even when analyzed in combination, the currently known AD risk factors explain only a fraction of the phenotypic variance[7], and, accordingly, only have limited applicability as early markers for disease onset and progression[7,10]. Second, most of the currently reported AD susceptibility genes were identified using classic case–control designs comparing clinically manifest dementia-stage AD vs. control individuals, typically lacking data on early-stage impairments (e.g., mild cognitive impairment [MCI]) and clinical follow-up to ascertain progression and eventually conversion to AD. Finally, while some studies have investigated the correlation between genetics and non-genetic biomarkers, this was hitherto typically done as bivariate assessments owing to the lack of a broad spectrum of biomarkers and imaging data in the *same* individuals. To overcome at least some of these shortcomings we generated genome-wide single-nucleotide polymorphism (SNP) genotyping data in the European Medical Information Framework Alzheimer's Disease Multimodal Biomarker Discovery (EMIF-AD MBD) sample[11]. This powerful and unique dataset allows to combine genomic data (and "-omics" data from other domains) with preclinical biomarker levels to eventually improve our ability for an early detection and prevention of AD. While similar to the Alzheimer's Disease Neuroimaging Initiative (ADNI) study[12] in various aspects, EMIF-AD MBD extends ADNI and scope in several important ways, e.g., in the breadth of the biomarker assessments as well as the availability of "-omics" data from various different domains in the same individuals (for more details see Bos et al.[11]).

In this report, we focus exclusively on the description of the results from genome-wide association analyses using various Aβ and tau-relevant outcomes available in EMIF-AD MBD. Specifically, we performed GWAS and polygenic risk score (PRS) assessments for more than a dozen binary and quantitative phenotypes derived from five measures of cerebrospinal fluid (CSF) Aβ and tau proteins in addition to using simple diagnostic status (i.e., AD, MCI, and control). Whenever available, we compare our findings using equivalent GWAS results from the ADNI dataset.

## Materials and methods

### Sample and phenotype description

Overall, the EMIF-AD MBD dataset comprises 1221 elderly individuals (years of age: mean = 67.9, SD = 8.3) with different cognitive diagnoses at baseline (NC = normal cognition; MCI = mild cognitive impairment; AD = AD-type dementia). In addition, Aβ status, cognitive test results and at least two of the following were available at baseline for analyses in all EMIF-AD MBD individuals: plasma ($n = 1189$), DNA ($n = 929$), magnetic resonance imaging (MRI; $n = 862$), or CSF ($n = 767$ individuals). Furthermore, clinical follow-up data were available for 759 individuals. The demographic information of the 16 outcome phenotypes (9 binary and 7 quantitative) of the EMIF-AD MBD dataset utilized in this paper is summarized in Table 1. Depending on the availability of the clinical records, each phenotype has different effective sample sizes. We categorized the phenotypes analyzed in this study into three main categories, i.e., "diagnosis", "amyloid protein assessment" and "tau protein assessment" (NB: the diagnostic criteria used here for AD are not incorporating biomarker status so that some AD cases were classified as "amyloid negative"). Details related to sample ascertainment and phenotype/biomarker collection in EMIF-AD MBD have been described previously[11] and are summarized for the relevant traits of this study in Supplementary Table 1. Whenever available, we attempted to validate EMIF-AD MBD findings in the independent ADNI dataset using identical or comparable phenotypes (this was possible for the two diagnostic groups, as well as for 6 amyloid- and 2 tau-related traits; Table 1). The local medical ethical committee in each participant recruitment center approved the study. Subjects had provided written informed consent at the time of inclusion in the cohort for use of data, samples and scans[11].

Replication data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET),

**Table 1 Overview of binary and quantitative traits available for genome-wide association study (GWAS) and polygenic risk score (PRS) analyses in EMIF-AD MBD and ADNI datasets.**

| | | | | EMIF-AD MBD (Discovery) | | | ADNI (Replication) | | |
|---|---|---|---|---|---|---|---|---|---|
| **Category** | **Variable type** | **Variable** | **Description** | **#sample_n** | **GWAS** | **PGS** | **Variable** | **#sample_n** | **GWAS** |
| Clinical diagnosis | Binary | AD vs. NC | Alzheimer disease (AD) vs. normal cognition (NC) | 545 (212 AD vs. 333 NC) | YES | YES | AD vs. NC | 303 (46 AD vs. 257 NC) | YES |
| | | MCI vs. NC | Mild cognitive impairment (MCI) vs. normal cognition (NC) | 659 (326 MCI vs. 333 NC) | YES | YES | MCI vs. NC | 705(448 MCI vs. 257 NC) | YES |
| Amyloid protein assessment | Binary | AMYLOIDstatus_ALL | Dichotomous amyloid classification variable across all diagnostic groups | 871 (455 abnormal vs. 416 normal) | YES | YES | AMYLOIDstatus | 618 (361 abnormal vs. 257 normal) | YES |
| | | AMYLOIDstatus_MCI | Dichotomous amyloid classification variable in MCI subjects | 326 (189 abnormal vs. 137 normal) | YES | YES | AMYLOIDstatus_MCI | 371 (232 abnormal vs. 139 normal) | YES |
| | | AMYLOIDstatus_NC | Dichotomous amyloid classification variable in NC subjects | 333 (77 abnormal vs. 256 normal) | YES | YES | AMYLOIDstatus_NC | 202 (88 abnormal vs. 114 normal) | YES |
| | | Central_CSF_ratiodich | Dichotomous variable based on ratio of central CSF amyloid-42/40 values | 677 (418 abnormal vs. 259 normal) | YES | YES | NA | NA | NA |
| | | Local_AB42_Abnormal | Dichotomous variable of local CSF amyloid-beta-42 values | 726 (392 abnormal vs. 334 normal) | YES | YES | AB42_abnormal | 578 (340 abnormal vs. 238 normal) | YES |
| | Quantitative | AB_Zscore | Z-score for amyloid pathology | 890 | YES | YES | AB_Zscore | 578 | YES |
| | | log_Central_CSF_AB42 | Log-transformed central CSF amyloid-beta-42 values | 677 | YES | YES | ABETA.Lumi.bl | 578 | YES |
| | | Central_CSF_AB38 | Central CSF Amyloid-beta-38 values | 675 | YES | YES | ABETA38.MSM.bl | 548 | YES |
| | | Central_CSF_AB40 | Central CSF Amyloid-beta-40 values | 677 | YES | YES | ABETA40.MSM.bl | 548 | YES |
| | | log_Central_CSF_AB4240ratio | Ratio of log-transformed central CSF amyloid-beta-42 vs. 40 values | 677 | YES | YES | NA | NA | NA |
| Tau protein assessment | Binary | Local_TTAU_Abnormal | Dichotomous variable of local CSF total-tau values | 724 (378 abnormal vs. 346 normal) | YES | YES | NA | NA | NA |
| | | Local_PTAU_Abnormal | Dichotomous variable of local CSF phospho-tau values | 726 (354 abnormal vs. 372 normal) | YES | YES | NA | NA | NA |
| | Quantitative | Ttau_ASSAY_Zscore | Z-score for CSF total-tau values | 723 | YES | YES | log_TTAU_Zscore | 571 | YES |
| | | Ptau_ASSAY_Zscore | Z-score for CSF phospho-tau values | 726 | YES | YES | log_PTAU_Zscore | 576 | YES |

other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild MCI and early AD. The ADNI participants utilized for our analyses originate from both ADNI1 and ADNIGo/2 and relate to those with available whole-genome sequencing (WGS; see below) data. Accordingly, we label the subset of ADNI participants utilized here as "ADNI-WGS".

### DNA extraction

Our laboratory at University of Lübeck, Germany, had access to 953 DNA samples from EMIF-AD MBD participants[11] for genetic (this paper) and epigenetic (DNA methylation profiling, m.s. in preparation) experiments. All participants had provided written consent to these experiments and institutional review board (IRB) approvals for the utilization of the DNA samples in the context of EMIF-AD MBD were obtained by the sample collection sites. For 805 participants, DNA was extracted locally at the collection sites. For 148 whole-blood samples DNA extraction was performed in our laboratory using the QIAamp® DNA Blood Mini Kit (QIAGEN GmbH, Hilden, Germany). Overall, this resulted in a total number of 953 DNA samples available for subsequent processing and analysis. Quality control (QC; by agarose gel electrophoresis, determination of A260/280 and A260/230 ratios, and PicoGreen quantification) resulted in 936 DNA samples of sufficient quality and quantity to attempt genome-wide SNP genotyping using the Infinium Global Screening Array (GSA) with Shared Custom Content (Illumina Inc.). GSA genotyping was performed at the Institute of Clinical and Medical Biology (UKSH, Campus-Kiel) on an iScan instrument (Illumina, Inc) following the manufacturer's recommendations. All 936 DNA samples passed post-experiment QC according to the manufacturer's instructions.

### Genotype imputation and quality control

Data processing was performed from raw intensity data (idat format) in GenomeStudio software (v2.0.4; Illumina, Inc.). We then used PLINK software (v1.9)[13] to perform pre-imputation QC and bcftools (v1.9)[14] to remove ambiguous SNPs, flipping and swapping alleles to align to human genome assembly GRCh37/hg19 before imputation. The QC'ed data (i.e., 931 samples and 498 589 SNPs) were then phased using SHAPEIT2 (v2.r837)[15] and imputed locally using Minimac3[16] based on a pre-compiled Haplotype Reference Consortium (HRC) reference panel (EGAD00001002729 including 39,131,578 SNPs from ~11 K individuals). Following post-imputation QC, we retained a total of 7,778,465 autosomal SNPs with minor allele frequency (MAF) ≥ 0.01 in 898 individuals of European ancestry for downstream association analysis. A full description of data processing and QC procedures is provided in the Supplementary Material.

### Classification of *APOE* genotypes

For all but 80 samples *APOE* genotype (i.e., for SNPs rs7412 [a.k.a. as "ε2-allele"] and rs429358 [a.k.a. "ε4-allele"]) was determined locally at the sample collection sites. To ensure that these prior genotypes correctly align to those resulting from genome-wide genotyping, local *APOE* genotypes were compared to those either inferred directly (i.e., rs7412) or indirectly (i.e., by imputation: rs429358) from GSA genotyping. These comparisons resulted in a total of 5 mismatches (~0.6%). In these 5 and the 80 samples without prior *APOE* genotype information, genotyping was determined manually in our laboratory using TaqMan assays (ThermoFisher Scientific, Foster City, CA) on a QuantStudio-12K-Flex system in 384-well format. TaqMan re-genotyping confirmed all five local genotype calls (which were used as genotypes in all subsequent analyses).

### Biochemical analyses of CSF biomarkers

CSF sampling and storage conditions have been described elsewhere[11]. CSF concentrations of Aβ38, Aβ40 and Aβ42 were measured using the V-PLEX Plus Aβ Peptide Panel 1 (6E10) Kit from Meso Scale Discovery (MSD, Rockville, MD). The measurements were performed at the Clinical Neurochemistry Laboratory in Gothenburg in one round of experiments, using one batch of kit reagents, by board-certified laboratory technicians, who were blinded to clinical data. For phosphorylated tau (Ptau) and total tau (Ttau), available data from the local cohorts were used. These were derived in clinical laboratory practice using INNOTEST ELISAs (Fujirebio, Ghent, Belgium), as previously described[17]. In the absence of CSF for new analyses of CSF Aβ proteins, we used local INNOTEST ELISA-derived CSF Aβ42 data to allow classifying as many subjects as possible as either Aβ-positive or -negative (see ref. [11] and below).

### GWAS and post-GWAS analyses

SNP-based association tests were performed using logistic regression models in mach2dat[18,19], for binary traits and linear regression models in mach2qtl[18,19], for quantitative traits. Association analyses utilized imputation-derived allele dosages as independent variables and were adjusted for sex, age at examination, and principle components (PC) 1 to 5 (using PLINK –pca to compute eigenvalues for up to 20 PCs; the number of PCs was then determined visual inspection of the scree plot). Diagnostic groups (coded as AD = 3, MCI = 2, controls = 1) were included as additional covariates in all analyses except for diagnostic outcome. QQ and Manhattan plots were constructed in R version 3.3.3 (https://www.r-project.org) using the "qqman" package[20]. The genomic inflation factor was calculated in R using the "GenABEL" package[21]. Statistical significance for the SNP-based

analyses was defined as $\alpha = 5E\text{-}08$, a widely used threshold that accounts for the approximate number of independent variants (~1 M) in European populations[22,23]. Post-GWAS, we used FUMA (http://fuma.ctglab.nl/)[24] to perform functional mapping and annotation of the genome-wide association results. This included calculating gene-based association statistics using MAGMA[25] using predefined sets of genes as implemented in FUMA. Statistical significance for the gene-based analyses was defined as $\alpha = 0.05/18720 = 2.671E\text{-}06$ based on the number of genes ($n = 18720$) utilized for these analyses, as suggested by FUMA[24].

### Polygenic risk score (PRS) analysis

PRS were calculated for each individual from the summary statistics of two partially overlapping AD case–control GWAS, i.e., the paper by Jansen et al.[7] including data from >380,000 individuals from the UK biobank, and a 2013 GWAS meta-analysis from the International Genomics of Alzheimer's Project (IGAP)[8]. Note that the genome-wide screening data of the IGAP study ("stage 1") was also included in the meta-analysis by Jansen et al. After removal of ambiguous SNPs (A/T and C/G) and filtering SNPs by MAF > 0.01 and imputation quality Rsq >0.8, PLINK 1.9 software[13] was used for linkage disequilibrium (LD) pruning and scoring for a variety of P-value thresholds (5E-08, 5E-06, 1E-04, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 1.00). The resulting PRSs were used as independent variable in the regression models adjusting for sex, age, and PC1 to PC5 as covariates. For the phenotypes not representing the diagnostic outcome, we also included diagnosis as additional covariate. For linear models, variance explained ($R^2$) was derived from comparing results from the full model (including PRS and covariates) vs. the null model (linear model with covariates only). Using a similar partitioning approach, we estimated the percent trait variance explained by our GWAS results for the five CSF traits. For logistic models, we calculated Nagelkerke's $r^2$ using the R package fmsb. A full description of these and all other statistical procedures is provided in the "Supplementary Methods".

### Validation analyses in ADNI

Whenever possible, we used whole-genome sequencing data from the ADNI cohort to assess replicability of the EMIF-AD MBD findings. The ADNI-WGS sample used here comprises 808 subjects with available whole-genome sequencing data (for more details on the generation of these data, see http://adni.loni.usc.edu/study-design/). For the analyses performed here, we only used unrelated subjects of European origin ($n = 751$). Variant-based filtering was performed based on minor allele count (MAC > 3), missingness rate (not more than 5 %) and Hardy-Weinberg equilibrium ($P > 1E\text{-}05$). We calculated principal components to account for population stratification based on an LD-pruned subset of common variants (MAF > 0.1). Association statistics were calculated using PLINK v2.0 using linear and logistic regression models (as appropriate), controlling for age, sex and four principal components as basic covariates in our models. For the phenotypes not representing the diagnostic outcome, we included diagnosis as an additional covariate. For more information on ADNI, please see Supplementary Material.

## Results

### GWAS on diagnostic outcomes

First, we performed GWAS for diagnostic outcomes, i.e., all cases diagnosed with AD ($n = 212$) and MCI ($n = 326$), against normal control subjects ($n = 333$). Comparing AD vs. controls showed the expected strong signals in the *APOE* region on chromosome 19q reaching genome-wide suggestive significance in the SNP-based analyses (best SNP rs429358: OR = 2.26, 95% CI = 1.66–3.07, $P = 2.68E\text{-}07$; Supplementary Fig. 1 and Supplementary Table 2) and genome-wide significant association in the gene-based tests (*APOC1* [$P = 3.24E\text{-}07$] and *APOE* [$P = 3.39E\text{-}07$]; Supplementary Fig. 1B and Supplementary Table 2). Interestingly, and in contrast to most other previous AD GWAS (e.g., Jansen et al.[7] and Lambert et al.[8]), the best-associated SNP (rs429358) in this region is the variant defining the "ε4" allele in the commonly used "ε/2/3/4" haplotype (the "ε2" allele is defined by rs7412). *APOE* was also the top-associated region in the ADNI dataset (Supplementary Table 2), as previously described[26]. Interestingly, in analyses comparing MCI vs. controls, the *APOE* region did not emerge as strongly associated (i.e., *P*-value for rs429358 = 0.17; Supplementary Fig. 2 and Supplementary Table 3). Instead, the best-associated SNP was rs153308 (OR = 0.51, 95% CI = 0.39–0.67, $P = 1.25E\text{-}06$; Supplementary Fig. 2A, Supplementary Table 3), located in an intergenic region on chromosome 5q13.3. However, neither this nor any of the other variants showing *P*-values <1E-05 showed any evidence of association in the ADNI dataset, thus they may—at least in part—not reflect genuine association signals.

### GWAS on dichotomous "amyloid classification"

In the remainder of our analyses we focus on AD-relevant CSF biomarker data as phenotypic outcomes of our GWAS analyses (see Table 1 for an overview of all traits analyzed). These analyses included a dichotomous "amyloid classification" (normal/abnormal) variable, representing a combination of CSF Aβ42/40 ratio, local CSF-Aβ42, and the standardized uptake value ratio (SUVR) on an amyloid PET scan (see methods section on "amyloid classification" in Bos et al.[11] for more details).

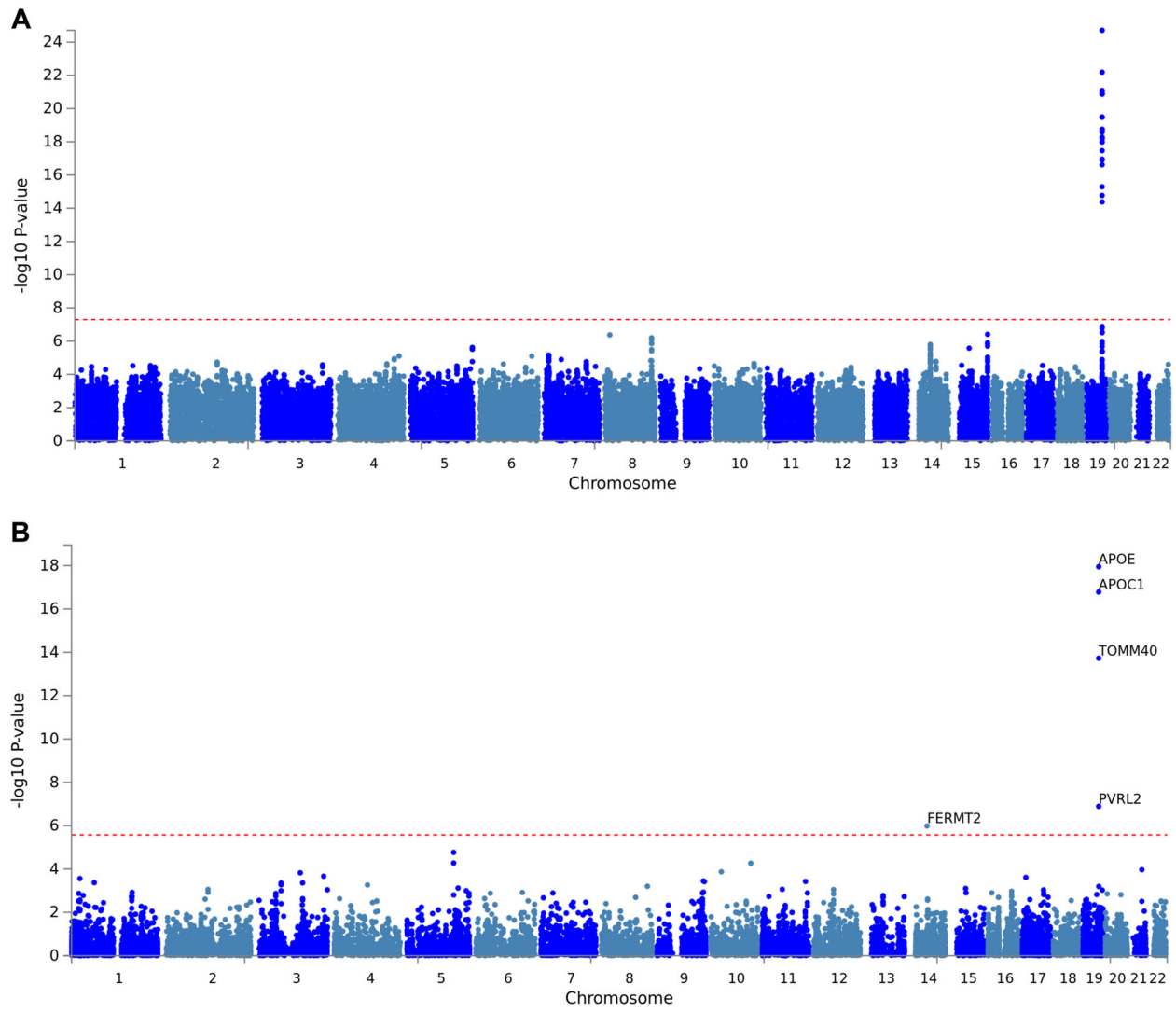


**Fig. 1 GWAS results using amyloid status in the EMIF-AD MBD dataset.** Manhattan plots of **A** SNP level and **B** gene-level genome-wide association results with the "amyloid classification" variable across all diagnostic groups ($n = 871$). All plots include gene assignments made with FUMA. Dotted red lines represent the threshold for genome-wide significance, i.e., $a = 5.0E\text{-}08$ for SNP-based (**A**) and $a = 2.671E\text{-}06$ for gene-based (**B**) analyses.

Using this amyloid variable (available for $n = 871$ individuals, of which $n = 455$ were classified as "abnormal" and $n = 416$ as "normal") across all diagnostic groups, we observed multiple genome-wide significant signals in the *APOE* region on chromosome 19 (Fig. 1). Similar to the GWAS on diagnostic outcome, the strongest association was observed with the *APOE* "ε4" allele (i.e., SNP rs429358: OR = 5.66, 95% CI = 4.09−7.84, $P = 1.95E\text{-}25$; Fig. 1A and Supplementary Table 4), which was also very strong in equivalent analyses in the ADNI dataset ($P = 4.15E\text{-}22$). Despite the consistency of the *APOE* findings, none of the other suggestive signals outside the *APOE* region replicated in ADNI (Supplementary Table 4). As expected, gene-based tests using MAGMA highlighted *APOE* (and neighboring loci) as the most significant gene(s) associated with the "amyloid classification" variable ($P = 1.13E\text{-}18$, Fig. 1B and Supplementary Table 4). In addition, a second locus emerged at genome-wide significance in the gene-based analyses, i.e., fermitin family homolog 2 (*FERMT2*) ($P = 1.03E\text{-}06$) located on chromosome 14q22.1. While this gene was originally reported to represent an AD risk locus[8], this finding was not replicated in the much larger GWAS by Jansen et al.[7]. Furthermore, gene-based tests in ADNI did not reveal evidence for association between markers in *FERMT2* and the "amyloid classification" phenotype ($P = 0.23$), suggesting that this gene is at best marginally involved in determining variance of this trait. Additional analyses using the "amyloid classification" variable limited to MCI and control individuals revealed a similar picture as in the

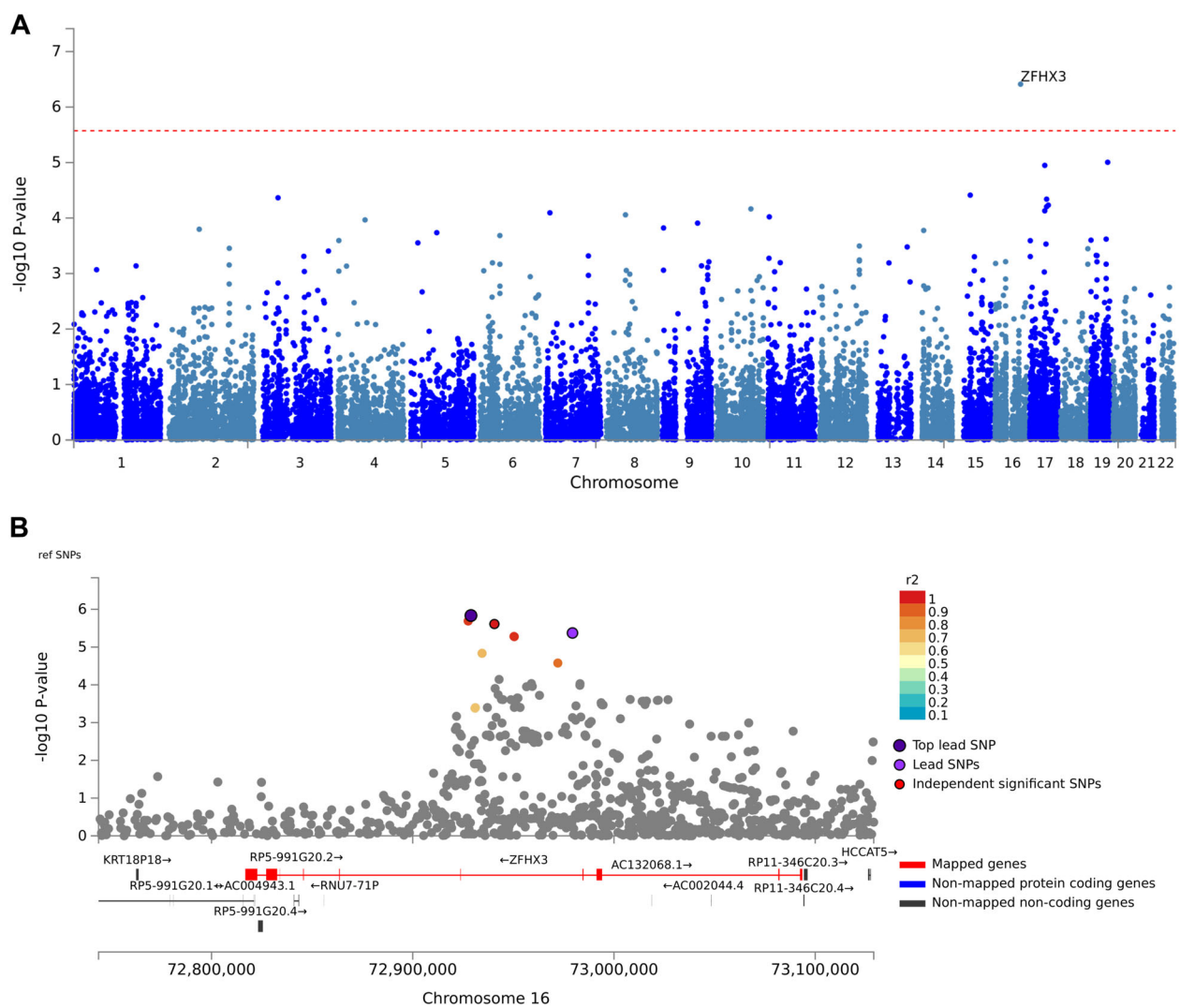


**Fig. 2 GWAS results using CSF-Aβ38 in the EMIF-AD MBD dataset.** Manhattan plots of **A** gene-level genome-wide association results with CSF-Aβ38 levels across all diagnostic groups ($n = 675$), **B** regional association results zoomed into a 450 kb region surrounding the *ZFHX3* gene on chromosome 16q22. All plots include gene assignments and linkage disequilibrium estimates made with FUMA. Dotted red line represents the threshold for genome-wide significance ($a = 2.671E-06$) for the gene-based analyses.

full dataset (see Supplementary Tables 5 and 6, respectively), i.e., the strongest association signals were observed for *APOE* (all replicated in equivalent analyses in ADNI). In addition, we identified a few suggestive non-*APOE* signals on other chromosomes, albeit none of these showed evidence for independent replication in ADNI.

### GWAS on additional dichotomous and continuous CSF amyloid variables

Overall, there were a total of two binary ("Central_CSF_ratiodich" and "Local_AB42_Abnormal") and five quantitative ("AB_Zscore", "Central_CSF_AB38", "Central_CSF_AB40", "log_Central_CSF_AB42" and "log_Central_CSF_AB4240ratio"; see Table 1) CSF-Aβ phenotypes available that were used as outcome variables in the GWAS. With the exception of "CSF-Aβ38" and "CSF-Aβ40" all showed strong and highly significant association with markers in the *APOE* region but no better than suggestive ($P < 1E-05$) signals in the remaining genome (see Supplementary Figs. 3–7 and Supplementary Tables 7–11). None of the non-*APOE*, suggestive signals replicated in the ADNI dataset for phenotypes with available data. GWAS results on "CSF-Aβ38" (Fig. 2, Supplementary Fig. 8 and Supplementary Table 12) and "CSF-Aβ40" (Supplementary Fig. 9 and Supplementary Table 13) showed highly similar GWAS results owing to the—well-established[27]—high correlation between both markers, which we also observed in this dataset (Pearson's

$r^2 = 0.96$, *P*-value < 2.2E-16 across all samples with available data). The most noteworthy finding emerging from these analyses was genome-wide significant association between CSF-Aβ40 and markers in the gene encoding zinc finger homeobox 3 (*ZFHX3*) in gene-based analyses (gene-based P-value = 7.48E-08; best SNP *P*-value = 7.29E-08; Supplementary Fig. 9 and Supplementary Table 13; similar results were obtained with CSF-Aβ38, Fig. 2). Despite the highly significant and consistent association between CSF-Aβ40 and CSF-Aβ38 levels and *ZFHX3*, this finding was not replicated in ADNI (gene-based *P*-value = 0.77; Supplementary Table 13). Genome-wide significant (*P* < 5E-08) and suggestive (*P* < 5E-06) SNPs explained 29-32% of the phenotypic variance in our dataset (Supplementary Table 14). However, we note that this number likely overestimates the true variance explained as GWAS and genome partitioning were performed in the same dataset.

### GWAS on dichotomous and continuous CSF tau variables

Similar to the GWAS analyses for CSF Aβ measures, there were several dichotomous and continuous measures of CSF tau available in the EMIF-AD MBD dataset (see Table 1, Supplementary Figs. 10–13, Supplementary Tables 15–18). Using CSF-Ttau levels as z-scored continuous outcome we identified genome-wide significant association with markers in geminin coiled-coil domain containing (*GMNC*) on chromosome 3q28 in gene-based analyses (*P* = 1.61E-06; Supplementary Fig. 11B, Supplementary Table 16). In the SNP-based analyses, markers in this gene showed genome-wide suggestive P-values ranging between 4.3E-06 and 9.5E-06 (Supplementary Table 16). These results showed marginal evidence for association in the ADNI cohort on the SNP level (for rs6444469 with *P*-value = 0.09), but not on the gene-level (*P*-value = 0.59; Supplementary Table 16). Interestingly, association between markers in *GMNC* and CSF-Ttau levels were previously described[28]. A follow-up study to the original report provided evidence for sex-specific differences at this locus (rs1393060 proximal to *GMNC* and in strong LD with rs6444469 [$r^2 = 1$]: *P*-value = 4.57E-10 in females compared to *P*-value = 0.03 in males), suggesting stronger effects in females[29]. In EMIF-AD MBD, we also observed a stronger association in females for *GMNC* at the gene and SNP level, respectively (gene: *P*-value = 4.09E-06 [in females] vs. *P*-value = 0.08 [in males]; SNP [for rs6444469]: *P*-value = 1.47E-04 [in females] vs. *P*-value = 8.91E-03 [in males]), hence, providing independent replication of the previous report. Similarly, the association results for rs6444469 and CSF-Ttau in ADNI were more pronounced in females (*P*-value = 0.013) than males (*P*-value = 0.93). The other available CSF tau measure in our GWAS was CSF-Ptau, which is known to strongly correlate with CSF-Ttau levels[30], a correlation, which we also observe in our data (Pearson $r^2 = 0.87$, *P*-value < 2.2E-16). Owing to this phenotypic correlation, the GWAS results for both variables also look quite similar, as expected (Supplementary Fig. 13B and Supplementary Table 18), and replicate the sex-specific difference. Genome-wide significant (*P* < 5E-08) and suggestive (*P* < 5E-06) SNPs explained about 28–34% of the phenotypic variance in our dataset (Supplementary Table 14). However, we note that this number likely overestimates the true variance explained as GWAS and genome partitioning were performed in the same dataset.

### Polygenic risk score (PRS) analyses on all outcome traits

In addition to testing all above-mentioned traits for SNPs and genes in the context of genome-wide analyses, we also computed association statistics with aggregated variant data in the form of PRS using AD case–control results from Lambert et al.[8] and Jansen et al.[7]. The aim was to assess the degree at which established AD-associated markers also show association with the "AD-related" (endo-) phenotypes analyzed here. Although a considerable amount of sample overlap exists across the two AD risk GWAS (i.e., both use the "stage I" GWAS data from the IGAP sample in their discovery phase), both studies use different analysis paradigms and different replication cohorts. Given that the final effective sample size used in Jansen et al. is nearly ten-times larger than that in Lambert, we hypothesized that the results and, accordingly PRS, derived from the larger study are more "precise" and will, therefore, show stronger association and explain more of the trait variance analyzed here. PRS-based results are summarized in Table 2, while full results can be found in Supplementary Table 19 (for PRS from Jansen et al.[7]) and 20 (for PRS from IGAP[8]). Overall, we observed significant PRS-based associations with many, but not all, traits analyzed in the EMIF-AD MBD dataset. For both PRS models, the best-associated trait was a diagnosis of AD and all measures involving CSF-Aβ42 levels. In contrast, no noteworthy associations were observed with CSF-Aβ38 and CSF-Aβ40 levels and, perhaps more interesting, not with either of the two available CSF-tau measures (CSF-Ttau and CSF-Ptau). Comparing the "performance" of both PRS against each other revealed that—against our expectation—the IGAP-based results tended to show the stronger statistical support (i.e., smaller P-values) and explained slightly more of the phenotypic variance (i.e., showed higher $r^2$) than the PRS derived from more recent and larger GWAS by Jansen et al. (Supplementary Table 19 vs. 20). The only exception being the case–control analyses on AD status, where the Jansen PRS outperformed that from IGAP (i.e., $r^2$ = 4.35%, *P*-value = 4.03E-06 vs. $r^2$ = 2.98%, *P*-value = 1.07E-04, respectively). Interestingly, the association of AD-based PRS with risk for MCI was minor in both

**Table 2 Summary of polygenic risk score (PRS) analyses using two *P*-value thresholds and two different GWAS datasets with and without markers in the *APOE* region.**

| | PRS constructed based on GWAS by Jansen et al.[7] | | | | | | | | PRS constructed based on GWAS by IGAP, 2013 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Including *APOE* | | | | Excluding *APOE* | | | | Including *APOE* | | | | Excluding *APOE* | | | |
| | S1($P < 5E-8$) | | S5($< 0.05$) | | S1($P < 5E-8$) | | S5($< 0.05$) | | S1($P < 5E-8$) | | S5($< 0.05$) | | S1($P < 5E-8$) | | S5($< 0.05$) | |
| Phenotype | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value | $r^2$ | *P*-value |
| AD | *3.09%* | *8E-05* | *2.89%* | *0.0002* | *1.22%* | *0.0126* | *0.99%* | *0.0244* | **2.98%** | **0.0001** | *1.69%* | *0.0034* | **0.78%** | **0.0443** | 0.10% | 0.4745 |
| MCI | 0.12% | 0.3767 | *0.77%* | *0.0249* | 0.07% | 0.4894 | *0.61%* | *0.0449* | 0.51% | 0.0682 | 0.13% | 0.3537 | 0.49% | 0.0743 | 0.00% | 0.8764 |
| AMYLOIDstatus | *1.38%* | *0.0004* | *1.01%* | *0.0025* | 0.38% | 0.0643 | 0.04% | 0.5599 | *3.82%* | *9E-09* | *2.95%* | *3E-07* | **1.09%** | **0.0018** | 0.34% | 0.0766 |
| Amyloid.MCI | *4.59%* | *0.0008* | *5.82%* | *0.0002* | 1.33% | 0.0684 | 0.87% | 0.1394 | **12.54%** | **1E-07** | *7.45%* | *2E-05* | **4.63%** | **0.0008** | 0.82% | 0.1513 |
| Amyloid.NC | 0.83% | 0.1637 | 0.23% | 0.4657 | 0.20% | 0.4928 | 0.01% | 0.8786 | *2.45%* | *0.0174* | *5.10%* | *0.0008* | 0.24% | 0.4592 | 1.28% | 0.0866 |
| AB_Zscore | *1.20%* | *0.0003* | *0.80%* | *0.0031* | 0.27% | 0.0871 | 0.01% | 0.7683 | *2.61%* | *7E-08* | *3.08%* | *4E-09* | **0.39%** | **0.0382** | *0.38%* | *0.04* |
| Central_CSF_ratiodich | **3.86%** | **2E-06** | *1.94%* | *0.0007* | **1.49%** | **0.0029** | 0.10% | 0.4414 | **5.87%** | **9E-09** | *3.52%* | *6E-06* | **1.58%** | **0.0022** | 0.22% | 0.2439 |
| Local_AB42_Abnormal | **3.68%** | **4E-06** | *1.13%* | *0.0098* | **1.65%** | **0.0019** | 0.02% | 0.7287 | **3.40%** | **1E-05** | *1.66%* | *0.0018* | **0.94%** | **0.0184** | 0.01% | 0.7873 |
| Central_CSF_AB38 | 0.21% | 0.2273 | 0.00% | 0.8587 | 0.32% | 0.1305 | 0.01% | 0.7838 | 0.00% | 0.8816 | 0.28% | 0.1614 | 0.03% | 0.6416 | 0.33% | 0.1279 |
| Central_CSF_AB40 | 0.24% | 0.1805 | 0.01% | 0.7974 | 0.28% | 0.1514 | 0.05% | 0.5298 | 0.05% | 0.5418 | 0.44% | 0.0718 | 0.07% | 0.4874 | 0.36% | 0.106 |
| log_Central_CSF_AB4240ratio | *2.06%* | *4E-05* | *1.68%* | *0.0002* | *0.49%* | *0.0453* | 0.07% | 0.4428 | **3.74%** | **3E-08** | *2.55%* | *5E-06* | **0.56%** | **0.0323** | 0.05% | 0.5324 |
| log_Central_CSF_AB42 | *1.52%* | *0.0005* | *0.52%* | *0.0445* | **0.56%** | **0.0373** | 0.01% | 0.8297 | *2.31%* | *2E-05* | *2.17%* | *4E-05* | **0.53%** | **0.0425** | 0.25% | 0.1605 |
| Local_PTAU_Abnormal | 0.31% | 0.1602 | 0.03% | 0.662 | 0.06% | 0.5386 | 0.01% | 0.8148 | **0.69%** | **0.0347** | 0.38% | 0.1192 | 0.19% | 0.2697 | 0.09% | 0.4424 |
| Local_TTAU_Abnormal | **0.93%** | **0.0107** | 0.01% | 0.7505 | **0.56%** | **0.0465** | 0.12% | 0.3625 | **0.82%** | **0.0166** | 0.27% | 0.1644 | 0.17% | 0.267 | 0.04% | 0.5771 |
| Ptau_ASSAY_Zscore | 0.34% | 0.0909 | 0.00% | 0.9833 | 0.09% | 0.3922 | 0.05% | 0.5191 | 0.37% | 0.0806 | 0.08% | 0.4258 | 0.01% | 0.7755 | 0.01% | 0.8237 |
| Ttau_ASSAY_Zscore | **0.51%** | **0.032** | 0.03% | 0.6082 | 0.23% | 0.1529 | 0.00% | 0.9666 | 0.09% | 0.3566 | 0.11% | 0.3269 | 0.03% | 0.5811 | 0.00% | 0.8676 |

Italicized values = nominally significant association.
$r^2$ = variance explained; bold font = largest $r^2$ (= most variance explained) for trait in question.
A full listing of results from these PRS analyses can be found in Supplementary righthandcolumns of Table 2 and bottom part of SupplementaryTable 19 (for Jansen et al. GWAS) and Supplementary Table 20 (for IGAP GWAS).

models ($r^2 = 0.83\%$, *P*-value = 0.02, and $r^2 = 0.65\%$, *P*-value = 0.039, for Jansen and IGAP, respectively).

To investigate the contribution of markers in the *APOE* region, we repeated all analyses excluding variants within 1 MB of *APOE* (chr19:45409039-45412650; right-hand columns of Table 2 and affected were the analyses on AD (strongest reduction in bottom part of Supplementary Tables 19 and 20). As expected, removal of *APOE* region markers from the PRS decreased the variance explained for most of the traits analyzed here, albeit to varying degrees. Most affected were the analyses on AD (strongest reduction in $r^2 = 73.8\%$ in analyses excluding *APOE* effects vs. the full model including *APOE* in IGAP) and essentially all CSF-Aβ42 related measures (strongest reduction in $r^2 = 88.9\%$ for trait "AB_Zscore" in IGAP) for both PRS models. Less affected by the removal of *APOE* were the analyses of CSF-tau species (strongest reduction in $r^2 = 29.0\%$ for CSF-Ttau in non-*APOE* vs. *APOE* models).

## Discussion

This is the first GWAS utilizing part of the wide collection of AD-relevant phenotypes and biomarkers available in the EMIF-AD MBD dataset. The phenotypes analyzed here related either to clinical diagnosis (i.e., AD or MCI) or to levels of CSF biomarkers revolving around various biochemical species of amyloid or tau proteins. While GWAS results have already been reported for some of the biomarkers analyzed here (e.g., in ADNI), ours are the first to combine genomic and biomarker data in the newly established EMIF-AD MBD dataset. The main findings of our study can be summarized in the following five points: (1) the most prominent genetic signals in analyses of either diagnostic outcome or phenotypes related to CSF-Aβ42 were observed with markers in or near *APOE*, which is in good agreement with equivalent analyses in ADNI and other datasets; (2) our analyses identified one novel association in analyses of CSF-Aβ38 and CSF-Aβ40 levels and DNA sequence variants in *ZFHX3* (a.k.a. *ATBF1* [AT-motif binding factor 1]), although these signals were not replicated in the ADNI dataset; (3) using CSF-tau species (i.e., Ttau and Ptau), we confirmed the previously described association with SNPs in *GMNC*, including the recently reported effect modification by sex at this locus; (4) PRS analyses revealed that AD risk SNPs are mostly associated with phenotypes related to CSF-Aβ42 but not CSF-Aβ38, CSF-Aβ40, and most notably CSF-tau; (5) exclusion of *APOE* from the PRS analyses suggest that non-*APOE* AD GWAS SNPs explain at most 2.5% of the phenotypic variance underlying a diagnosis of AD in this dataset. Collectively, these results implicate that the genetic architecture underlying many traits relevant for AD research in EMIF-AD MBD compare well to other datasets of European descent paving the way for future genomic discoveries with additional phenotypes available in this unique cohort[11].

Despite these promising first results, our study is potentially confined by a number of possible limitations: First and foremost, despite the breadth of available phenotype data, the overall sample size of the EMIF-AD MBD dataset is comparatively small and, as a result, may not be adequately powered to detect genetic variants exerting smaller effects. To a degree, this limitation is alleviated by the fact that many available outcome phenotypes are of a quantitative nature, which are more powerful than analyses of dichotomous traits (e.g., disease risk). Second, many of the phenotypes available in EMIF-AD MBD are not currently ascertained in other, independent datasets (such as ADNI), making independent replication of any novel findings difficult. This situation can be expected to improve somewhat once the phenotypic breadth in ADNI and other cohorts is extended. Still, until independent replication is available, novel GWAS findings from EMIF-AD MBD must be interpreted with caution. This includes the putative association between CSF-Aβ38 and CSF-Aβ40 and markers in *ZFHX3*, which were highly significant and consistent in EMIF-AD MBD but not replicated in ADNI. Until more independent data on these outcome traits are available, the *ZFHX3* association should be considered "provisional". In this context it is comforting, however, that many well-established genetics findings (such as the association between *APOE* and measures of CSF-Aβ or *GMNC* and CSF-tau) were reproduced in EMIF-AD MBD. Third, while the genome-wide SNP genotype data was generated in one run of consecutive experiments in one laboratory, the same is not true for the phenotype measurements, which were performed locally in each of the 11 participating sites. In some instances, the compiled phenotype data are not based on the same biochemical assays across sites for some variables, e.g., measurements of tau protein. While the EMIF-AD MBD phenotype team went to great lengths to alleviate this potential problem by normalizing variables for each center (see Bos et al.[11] for more details), the possibility of artefactual findings owing to phenotypic heterogeneity remains. Finally, as described in the overall cohort description manuscript, the EMIF-AD MBD dataset is not designed to be "representative" of the general population but was assembled with the aim to achieve approximately equal proportions of amyloid+ vs. amyloid- individuals in all three diagnostic subgroups. While this ascertainment strategy does not invalidate our GWAS results per se, they may not be generalizable to the population as a whole. However, this limitation may affect any study with clinically ascertained participants and, thus, applies to most previously published GWAS in the field, including those performed in ADNI.

In conclusion, our first-wave of GWAS analyses in the EMIF-AD MBD dataset provides a first important step in a series of additional genome-wide and epigenome-wide (using DNA methylation profiles) association analyses in this valuable and unique cohort.

**Acknowledgements**

The present study was conducted as part of the EMIF-AD MBD project, which has received support from the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement n° 115372, the resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2007–2013) and EFPIA companies' in kind contribution. Parts of this study were made possible through support from the German Research Foundation (DFG grant FOR2488: Main support by subproject "INF-GDAC" BE2287/7-1 to LB). R.V. acknowledges the support by the Stichting Alzheimer Onderzoek (#13007, #11020, #2017-032) and the Flemish Government (VIND IWT 135043). H.Z. is a Wallenberg Academy Fellow supported by grants from the Swedish Research Council (#2018-02532), the European Research Council (#681712) and Swedish State Support for Clinical Research (#ALFGBG-720931). S. J.B.V. received funding from the Innovative Medicines Initiative 2 Joint Undertaking under ROADMAP grant agreement No. 116020 and from ZonMw during the conduct of this study. No conflict of interest exists. Research at VIB-UAntwerp was in part supported by the University of Antwerp Research Fund. The research was supported by ALF clinical grants from Region Västra Götaland to A.W. and to P.K. We acknowledge the assistance of Ellen De Roeck, Naomi De Roeck, and Hanne Struyfs (UAntwerp) with data collection. The Lausanne study was funded by a grant from the Swiss National Research Foundation (SNF 320030_141179) to J.P. We thank Mrs. Tanja Wesse and Mrs. Sanaz Sedghpour Sabet at the Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany for technical assistance with the GSA genotyping. The LIGA team acknowledges computational support from the OMICS compute cluster at the University of Lübeck. The computations in the ADNI cohort were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. For ADNI: Data was used for this project of which collection and sharing 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 NeuroImaging at the University of Southern California.

**Author details**

[1]Lübeck Interdisciplinary Platform for Genome Analytics (LIGA), Institutes of Neurogenetics and Cardiogenetics, University of Lübeck, Lübeck, Germany. [2]Genetics and Aging Unit and McCance Center for Brain Health, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [3]Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Alzheimer Centrum Limburg, Maastricht University, Maastricht, The Netherlands. [4]Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, The Netherlands. [5]Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, The Netherlands. [6]Institute of Neuroscience and Physiology, Department of Psychiatry and Neurochemistry, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [7]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Mölndal, Sweden. [8]Laboratory for Cognitive Neurology, Department of Neurosciences, KU Leuven, Leuven, Belgium. [9]Neurology Service, University Hospital Leuven, Leuven, Belgium. [10]Laboratory for Complex Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. [11]Neurochemistry Laboratory, Department of Clinical Chemistry, Amsterdam Neuroscience, Amsterdam University Medical Centers, Vrije Universiteit, Amsterdam, The Netherlands. [12]Department of Biomedical Sciences, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium. [13]Department of Neurology and Center for Neurosciences, UZ Brussel and Vrije Universiteit Brussel (VUB), Brussels, Belgium. [14]University of Geneva, Geneva, Switzerland. [15]IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy. [16]AIX Marseille University, INS, Ap-hm, Marseille, France. [17]Neurosciences Therapeutic Area, GlaxoSmithKline R&D, Stevenage, UK. [18]University of Lille, Inserm, CHU Lille, Lille, France. [19]Alzheimer's disease and other cognitive disorders unit, Hospital Clinic I Universitari, Barcelona, Spain. [20]Department of Neuropathology, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK. [21]Memory Unit, Neurology Department, Hospital de Sant Pau, Barcelona and Centro de Investigación Biomédica en Red en enfermedades Neurodegenerativas (CIBERNED), Madrid, Spain. [22]Geriatric Psychiatry, Department of Mental Health and Psychiatry, Geneva University Hospitals, Geneva, Switzerland. [23]Department of Psychiatry, University Hospital of Lausanne, Lausanne, Switzerland. [24]Department of Neurology, Center for Research and Advanced Therapies, CITA-Alzheimer Foundation, San Sebastian, Spain. [25]Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [26]NIHR BioResource Centre Maudsley, NIHR Maudsley Biomedical Research Centre (BRC) at South London and Maudsley NHS Foundation Trust (SLaM) & Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK. [27]Health Data Research UK London, University College London, 222 Euston Road, London, UK. [28]Institute of Health Informatics, University College London, 222 Euston Road, London, UK. [29]The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, 222 Euston Road, London, UK. [30]Steno Diabetes Center, Copenhagen, Denmark. [31]Institute of Pharmaceutical Sciences, King's College London, London, UK. [32]Neurodegenerative Brain Diseases Group, Center for Molecular Neurology, VIB, Antwerp, Belgium. [33]Department of Biomedical Sciences, University of Antwerp, Antwerp, Belgium. [34]Alzheimer Center and Department of Neurology, Amsterdam University Medical Centers, Amsterdam Neuroscience, Amsterdam, The Netherlands. [35]Department of Radiology and Nuclear Medicine, Amsterdam University Medical Centers, Amsterdam Neuroscience, Amsterdam, The Netherlands. [36]Institutes of Neurology and Healthcare Engineering, University College London, London, UK. [37]Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, Queen Square, London, UK. [38]UK Dementia Research Institute at UCL, London, UK. [39]Department of Psychiatry, University of Oxford, Oxford, UK. [40]Reference Center for Biological Markers of Dementia (BIODEM), Institute Born-Bunge, University of Antwerp, Antwerp, Belgium. [41]Translational Medicine Neuroscience, UCB Biopharma SPRL, Braine l'Alleud, Belgium. [42]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. [43]Department of Psychology, University of Oslo, Oslo, Norway

**Author contributions**

S.H. performed all the analyses and interpretation on data and wrote the manuscript. D.P. and R.T. contributed to replication in ADNI. V.D. was responsible for EMIF-AD MBD DNA sample preparation and extraction. F.K. was responsible for EMIF-AD MBD genotypic data QC and imputation. I.B., S.V., B.T., and P.J. coordinated the collection and harmonization of phenotypes and biosamples in EMIF-AD MBD and helped identifying equivalent phenotypes from the ADNI catalog. A.F. and M.W. supervised the genotyping experiments. U.A., K.B., and H.Z. performed CSF biomarker measurements and took part in cut-point determinations. K.S. and C.V.B. contributed to genetic characterization of samples and design of the genomics studies in EMIF-AD MBD, and critically revised the manuscript drafts. R.V., S.G., J.S., and I.C. contributed to sample and data collection. E.N. and I.S. were responsible for data collection of the Antwerp cohort. A.W. and P.K. contributed to data and sample collection/handling of the Gothenburg MCI Study, Gothenburg, Sweden. J.S., P.J.V., and S.L. are leads for the EMIF-AD MBD; designed and managed the platform. L.B. designed and supervised the genomics portion of the EMIF-AD MBD project and co-wrote all drafts of the manuscript. All authors

critically revised all manuscripts drafts, read, and approved the final manuscript.

### Data availability

GWAS summary statistics for the top (*P*-value < 1E-05) results are listed in the Supplementary Tables. Full GWAS summary statistics are available from the authors upon request. Clinical data and genome-wide genotyping data are stored on an online data platform using the "tranSMART" data warehouse framework. Access to the genome-wide genotyping data can be requested from the corresponding author of this study who will forward each request to the EMIF-AD data access team.

### Code availability

All scripts used to generate the primary GWAS and PRS analyses are available from the authors upon request.

### Conflict of interest

The authors declare that they have no conflict of interest.

### Publisher's note

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

**Supplementary Information** accompanies this paper at (https://doi.org/10.1038/s41398-020-01074-z).

Received: 6 March 2020 Revised: 23 April 2020 Accepted: 18 May 2020
Published online: 22 November 2020

### References

1. Sperling, R., Mormino, E. & Johnson, K. The evolution of preclinical Alzheimer's disease: implications for prevention trials. *Neuron* **84**, 608–622 (2014).
2. Mattsson, N. et al. Revolutionizing Alzheimer's disease and clinical trials through biomarkers. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* **1**, 412–419 (2015).
3. Tanzi, R. E. & Bertram, L. Twenty years of the alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell* **120**, 545–555 (2005).
4. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168 (2006).
5. Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* **39**, 17–23 (2007).
6. Kim, J., Basak, J. M. & Holtzman, D. M. The role of apolipoprotein E in Alzheimer's disease. *Neuron* **63**, 287–303 (2009).
7. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
8. Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
9. Bertram, L. et al. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. *Am. J. Hum. Genet.* **83**, 623–632 (2008).
10. Stocker, H., Möllers, T., Perna, L. & Brenner, H. The genetic risk of Alzheimer's disease beyond APOE ε4: systematic review of Alzheimer's genetic risk scores. *Transl. Psychiatry* **8**, 166 (2018).
11. Bos, I. et al. The EMIF-AD multimodal biomarker discovery study: design, methods and cohort characteristics. *Alzheimers Res. Ther.* **10**, 64 (2018).
12. Mueller, S. G. et al. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin. N. Am.* **15**, 869–877 (2005).
13. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
14. Danecek, P., Schiffels, S. & Durbin, R. *Multiallelic Calling Model In bcftools (-m)*. https://samtools.github.io/bcftools/.
15. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
16. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
17. Bos, I. et al. Cerebrospinal fluid biomarkers of neurodegeneration, synaptic integrity, and astroglial activation across the clinical Alzheimer's disease spectrum. *Alzheimer's Dement.* **15**, 644–654 (2019).
18. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
19. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
20. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Soft.*, **3**, 731, https://doi.org/10.21105/joss.00731 (2018).
21. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
22. T. I. H. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
23. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
24. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
25. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
26. Apostolova, L. G. et al. Associations of the top 20 Alzheimer disease risk variants with brain amyloidosis supplemental content. *JAMA Neurol.* **75**, 328–341 (2018).
27. Schoonenboom, N. S. et al. Amyloid β 38, 40, and 42 species in cerebrospinal fluid: More of the same?. *Ann. Neurol.* **58**, 139–142 (2005).
28. Cruchaga, C. et al. GWAS of cerebrospinal fluid tau levels identifies risk variants for Alzheimer's disease. *Neuron* **78**, 256–268 (2013).
29. Deming, Y. et al. Sex-specific genetic predictors of Alzheimer's disease biomarkers. *Acta Neuropathol.* **136**, 857–872 (2018).
30. Mulder, C. et al. Amyloid-beta(1-42), total tau, and phosphorylated tau as cerebrospinal fluid biomarkers for the diagnosis of Alzheimer disease. *Clin. Chem.* **56**, 248–253 (2010).