bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Title

A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease

Authors

Yongjin Park[+,1,2], Abhishek K Sarkar[+,3], Liang He[+,1,2], Jose Davila-Velderrain[1,2], Philip L De Jager[2,4], Manolis Kellis[1,2]

+: equal contribution.

1: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

2: Broad Institute of MIT and Harvard, Cambridge, MA, USA

3: Department of Human Genetics, University of Chicago, Chicago, IL, USA

4: Department of Neurology, Columbia University Medical Center, New York, NY, USA

MK: `manoli@mit.edu`

Abstract

Characterizing the intermediate phenotypes, such as gene expression, that mediate genetic effects on complex diseases is a fundamental problem in human genetics. Existing methods utilize genotypic data and summary statistics to identify putative disease genes, but cannot distinguish pleiotropy from causal mediation and are limited by overly strong assumptions about the data. To overcome these limitations, we develop Causal Multivariate Mediation within Extended Linkage disequilibrium (CaMMEL), a novel Bayesian inference framework to jointly model multiple mediated and unmediated effects relying only on summary statistics. We show in simulation that CaMMEL accurately distinguishes between mediating and pleiotropic genes unlike existing methods. We applied CaMMEL to Alzheimer's disease (AD) and found 206 causal genes in sub-threshold loci ($p < 10^{-4}$). We prioritized 21 genes which mediate at least 5% of local genetic variance, disrupting innate immune pathways in AD.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that evolves over decades and involves multiple molecular processes including accumulation of amyloid beta[1], propagation of neurofibrillary tangles in the brain, and ultimately cognitive decline[2]. However, the precise gene-regulatory mechanisms and molecular pathways involved in AD progression remain to be elucidated. Case-control studies of differential gene expression in large postmortem brain cohorts have identified multiple genes and regulatory elements associated by AD[3–5]. The gene regulatory networks of these differentially-expressed genes have revealed promising associated genes for AD[3,4], However, interpretation of case-control differential expression is challenging because it requires accounting for reverse causation of disease on gene expression, biological confounding, and technical confounding.

Large-scale meta analysis of genome-wide association studies (GWAS) for AD have identified 21 loci that influence AD susceptibility[6]. Unlike case-control differential expression, GWAS can in principle reveal causal relationships, because there is no possibility of reverse causation: late-onset complex disorders cannot change the genotypes at common genetic variants. However, the fundamental challenge in interpreting GWAS is that 90% of GWAS tag SNPs lie in noncoding regions[7], making it difficult to identify the target genes and the causal mechanisms with relevant cell type information. Causal genes are not necessarily closest to the GWAS tag SNP, but may instead be distal to the causal genetic variant and linked by long-ranged chromatin interactions[8–10].

Mendelian randomization[11,12] resolves causal directions by using genetic variants as instrumental variables (IV) in causal inference. However, MR assumes no unmediated effecs on phenotype, no horizontal pleiotropy[11,13] (the SNP affects both genes and phenotypes), and no LD-level linkage (two causal SNPs affecting genes and phenotypes in LD)[14]. A recent meta-analysis method, MR-Egger, partly relaxes these assumptions by modeling unmediated effects as a bias in regression of GWAS effect sizes on molecular QTL effect sizes[15]. However, it still assumes that genetic variants are not in LD (to perform meta-analysis) and that the estimated effect sizes reflect the true effect sizes of the variants (rather than being inflated by LD).

Recent methods for transcriptome-wide association studies (TWAS) have made important advances in identifying genes which could be causal. Unlike MR, TWAS aggregates information of multiple SNPs in LD to find genes whose *cis*-regulatory variants have correlated effect sizes for both gene expression and downstream phenotypes[16–18]. However, TWAS methods are fundamentally limited because they cannot distinguish between causal mediation, pleiotropy, linkage between causal variants, and reverse causation (Fig 1b), which could lead to inflated false positives as pleiotropy is quite prevalent in genetics[14]. Moreover, TWAS often finds multiple genes within a locus due to statistical correlations rather than independent biological effects because analysis is performed one gene at a time, ignoring gene-gene correlations[19].

Here, we present Causal Multivariate Mediation within Extended Linkage disequilibrium (CaMMEL), a new method for causal mediation analysis to find target genes from large-scale GWAS and molecular QTL trait association summary statistics (Fig 1a). With CaMMEL we address three aspects of the causal gene inference problem. First, CaMMEL leverages multiple SNPs in *cis*-regulatory window to account for LD. Second, CaMMEL explicitly models mediation effect of multiple genes in close proximity to select a sparse set of causal genes that explains away non-causal correlated genes. Finally, CaMMEL models mediation effects while adjusting for unmediated effects. In simulation we found that CaMMEL correctly distinguishes mediating genes from pleiotropic genes, achieving higher statistical power than TWAS[17,18] and Mendelian randomization with Egger regression (MR)[15].

We applied CaMMEL to identify mediating genes for AD, combining GWAS of 74,046 individuals (25,580 cases and 48,466 controls)[6] and brain gene expression of 356 individuals from the Religious Orders Study and the Memory and Aging Project (ROSMAP)[4,20]. We found 774 protein-coding genes with significant non-zero effect (FDR < 10^{-4}), of which 206 are located in proximity of subthreshold GWAS SNPs ($p < 10^{-4}$; Supp Tab 1), of which 66 explain at least 5% of local genetic variance. We further discussed 21 genes in subthreshold / GWAS loci that explain 5% of local variance, which includes the peroxisome regulator and cytochrome complex genes (*RHOBTB1*[21–23], *CYP27C1* and

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CYP39A1) and several microglial genes (*LRRC23*[24,25], *ELMO1*[26–28], *RGS17*[29], *CNTFR*[30]) which support functional roles of innate immune response in AD progression[2,31].

Results

Overview of causal mediation analysis with multiple genes and SNPs

The goal of causal mediation analysis is to disentangle mediated (indirect) and unmediated (direct) effects from total effects between exposure and outcome variables[32]. Here, we define the mediated effect as causal transcriptomic regulatory mechanism of the available eQTL data, and broadly define the unmediated effects to include non-causal horizontal pleiotropy[11], LD-level linkage[14], and unidentified mediated effects of other regulatory mechanisms. Unlike conventional approaches, we assume phenotype data (AD) are generated with multiple mediators[33,34] with coefficients $\boldsymbol{\beta}$ and the unmediated genetic effect with multivariate effect size $\boldsymbol{\gamma}$; expression mediators are generated from multivariate eQTL effects $\boldsymbol{\alpha}$ with unwanted reverse causation effect $\boldsymbol{\delta}$ (Fig 1c).

A standard method is to perform regressions in two stages[33]: First, estimate the multivariate regression of the mediators (genes) on the genotypes adjusting for reverse causation and other types of mediator-exposure confounding; then estimate the multivariate regression of phenotype on imputed mediators (potential outcome) with putative unmediated genetic effects. We can independently correct for the confounding effects in gene expression matrix by half-sibling regression (Online methods)[35], but the estimation of mediation effect sizes is fundamentally challenging because the mediator, e.g., gene expression, is usually not measured in the GWAS cohort, and imputation is not possible because the individual level genotypes required to impute the mediator are not available.

Model specification of summary-based causal mediation analysis

CaMMEL explicitly models the probabilities of multivariate mediation effects and unmediated effects on traits by reformulating the multivariate mediation models into the equivalent summary statistic-based model based on Regression with Summary Statistics (RSS)[36]. The key idea of RSS is to reformulate a individual-level multivariate regression to a generative model of the observed univariate effect size vector $\hat{\boldsymbol{\theta}}$ with the multivariate mean vector $\boldsymbol{\theta}$ and the covariance matrix estimated from LD matrix R (Online methods; eq. 5). We developed an efficient approximate inference algorithm for the RSS model and validated that it performed competitively with regression models on individual-level genotype data such as Bayesian Sparse Linear Mixed Model[37] (Supp Fig 3).

In CaMMEL, we model the total multivariate effect size vector $\boldsymbol{\theta}$ by decomposing it into a linear combination of multivariate eQTL effect sizes $\boldsymbol{\alpha}_k$ for each gene k with the coefficients β_k, and a multivariate effect size vector $\boldsymbol{\gamma}$ for the unmediated genetic effects:

$$\boldsymbol{\theta} = \sum_{k=1}^{K} \boldsymbol{\alpha}_k \beta_k + \boldsymbol{\gamma}. \tag{1}$$

The true multivariate eQTL effects, $\boldsymbol{\alpha}_k$, are *a priori* unknown, but can be inferred from the univariate eQTL effect sizes using RSS. In CaMMEL, we directly incorporate the RSS model for the univariate eQTL effect size vectors:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\sum_{k=1}^{K} \hat{\boldsymbol{\alpha}}_k \beta_k + SRS^{-1}\boldsymbol{\gamma}, SRS\right) \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where S is a diagonal matrix of standard errors (Online methods). We assume the spike-and-slab prior on the vector of mediation effect sizes $\boldsymbol{\beta}$ in order to find a sparse set of non-zero effects from the correlated mediators[38–41], but we assume a normal prior on direct effects $\boldsymbol{\gamma}$ to prevent the background unmediated effects from completely diminishing to zero.

CaMMEL generalizes the MR and TWAS methods in theory

To gain mathematical insights into the CaMMEL model, we derived the coordinate-wise maximum likelihood estimate (MLE) of mediation coefficient on gene k, $\hat{\beta}_k$ with estimated standard error $\hat{\tau}_k$ using the delta method[42]:

$$\hat{\beta}_k^{\text{MLE}} = \frac{\hat{\boldsymbol{\alpha}}_k^\top S^{-1} R^{-1} S^{-1} \hat{\boldsymbol{\theta}}}{(\hat{\tau}_k^{\text{MLE}})^{-2}} - \frac{\hat{\boldsymbol{\alpha}}_k^\top S^{-2} \boldsymbol{\gamma}}{(\hat{\tau}_k^{\text{MLE}})^{-2}} - \sum_{l \neq k} \beta_l \frac{\left(\hat{\boldsymbol{\alpha}}_k^\top S^{-1} R^{-1} S^{-1} \hat{\boldsymbol{\alpha}}_l\right)}{(\hat{\tau}_k^{\text{MLE}})^{-2}} \tag{3}$$

where estimated standard error $\hat{\tau}_k^{\text{MLE}} = \left(\hat{\boldsymbol{\alpha}}_k^\top S^{-1} R^{-1} S^{-1} \hat{\boldsymbol{\alpha}}_k\right)^{-1/2}$. We describe technical details and the derivation in the supplementary texts.

With a large sample size and consistent minor allele frequency between GWAS and eQTL cohorts, we can interpret the first term as covariance between GWAS trait and gene k, the second term as covariance between unmediated effect and gene k, and the third term as summation of influence from all the other genes. Under the asymptotic normality, the Wald test statistic on β_k converges to the standard normal distribution, $\hat{\beta}_k^{\text{MLE}} / \hat{\tau}^{\text{MLE}} \sim \mathcal{N}(0, 1)$.

TWAS[17] is a special case of CaMMEL where there is no directional bias in unmediated effect, $\mathbb{E}[\gamma] = 0$, and there there is no directional bias in mediated effects, $\mathbb{E}[\beta_l] = 0$ for all $l \neq k$. MR[15] is a special case of CaMMEL where there is no LD, $R = I$, no directional bias in mediated effects, but there could be directional bias of unmediated effects (which is modeled as a scalar parameter). As a consequence, TWAS and MR will only accurately estimate causal effect sizes if these conditions hold.

CaMMEL performs well in simulations and correctly controls unmediated effect

We simulated gene expression data and Gaussian phenotypes using genotypes of 1,709 individuals on 66 real extended loci (Online methods). We compared the performance of CaMMEL with TWAS[18] and the multi-SNP MR method[15]. Here, we show simulation results with two causal genes and two causal QTLs per gene as a representative example (Supp Fig 1), but we found similar trends with different number of causal genes and eQTLs per gene. When simulating with unmediated effects, we found CaMMEL achieved highest power and accuracy compared to TWAS and MR. At fixed FDR 1%, CaMMEL achieves 60% power to detect mediating genes when mediation explains 20% of genetic variance, and achieves nearly 50% power when mediation explains 10% of genetic variance (Supp Fig 1a). To investigate the robustness of methods against infused horizontal pleiotropy, we computed the percentage of falsely discovered genes at the threshold where each method achieved best precision and recall (maximum F1 score) in causal gene discovery (Supp Fig 1b). We found almost no evidence of spuriously associated genes in the discoveries made by CaMMEL, maintaining the fraction of false discoveries below 1%.

In contrast, TWAS, the statistical power and accuracy of TWAS degrade as we increase the level of unmediated effect. As expected by our analytical result, we found that TWAS had comparable power and AUPRC to CaMMEL only in the absence of unmediated effects (Supp Fig 1a). We confirmed weakened power of TWAS occurred when it confuses mediation with unmediated associated effect (Supp Fig 1b). We found unmediated associated genes are more frequently included in the top TWAS discoveries (5-10%) than CaMMEL (5%) and MR (2%). MR had essentially zero power in all of the simulated scenarios because the SNPs included in the model violate the independence assumption of the method (Supp Fig 1a).

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Transcriptome-wide mediation analysis in AD

We applied CaMMEL to AD GWAS and postmortem brain gene expression from the ROSMAP project to detect causal mediating genes in AD. We analyzed 2,077 independent, extended loci containing at least 100 well-imputed SNPs in the GWAS and ROSMAP. We allowed genes to span multiple LD blocks (based on SNPs within 1 Mb of the gene body) and empirically calibrated the null distribution of mediation by parametric bootstrap[43–46]. We controlled false discovery rate of the multiple hypothesis testing problem, estimating prior probability of alternative hypotheses from data[47]. Here, we defined *cis*-regulatory eQTLs and target genes with p-value < 0.05. We only considered genes with at least one eQTL after the p-value cutoff to avoid weak instrument bias in MR[48]. We included all GWAS SNPs in the background unmediated effect model if they appeared in the imputed genotype matrix, thus providing a conservative estimate of mediation effect, by allowing unmediated effect to explain away mediated effect.

We found 774 unique protein-coding genes with significant non-zero mediation effects (FDR < 10^{-4}). Of them, we found 43 are located in genome-wide significant loci, 163 in subthreshold loci (p < 10^{-4}; Fig 2a), and 568 in weakly associated regions (p > 10^{-4}). Of the 43 genes in GWAS loci, in 9 cases we found the lead SNPs from GWAS agree with cis-eQTL effects, indicating the primary genetic effects are found in the postmortem brain tissues. In the remaining cases strongest effects may exist in different tissues and may only become visible in single cell eQTL data.

We found on average the significant 774 genes each explain more than 2% of local genetically-driven phenotypic variance (proportion of variance explained, PVE). This proportion was similar for the 206 genes in the subthreshold/GWAS loci. However, the remaining genes explain essentially zero variance (Fig 2b). We found no clear evidence of correlation between GWAS significance with the proportion of local variance explained by mediation effects (Fig 2c) or the posterior probability of non-zero mediation effects (Fig 2d). Our results can be interpreted through omnigenic perspective[49] that a large number of causal genes accounting for weak polygenic effects do not yet reach genome-wide significance with current cohort size, suggesting causal variants of mediation genes located in the subthreshold regions can eventually reach genome-wide significance level as we increase sample size.

CaMMEL recapitulates known Alzheimer's disease genes

The 774 identified mediating genes include several known AD genes that are significant mediators and co-localized in genome-wide significant loci (p < 5e-8). *COPS6* accounts for nearly 9% of the variance (Fig 3a) and the overall eQTL effect sizes are negatively correlated with AD susceptibility (-0.21 $\pm$ 0.012; Tab 1). *COPS6* constitutes the *COP9* signalosome (CSN) complex, and the biological function of CSN was explored regarding innate immunity. The CSN complex is evolutionary conserved, and plays a key role in plant innate immunity[50]. Recently conditional knock-out experiments on the other subunit *COPS5* conducted in macrophages show highly correlated activities with anti-inflammatory pathways[51].

CLU explains nearly 6% of the local variance in the genome-wide significant locus (Fig 3b), and its mediation effects are positively correlated with AD susceptibility (mediation coefficient 0.18 $\pm$ 0.014; Tab 1). It is shown that *CLU* (clustrin) expression is elevated in AD brain and directly interacts with amyloid beta plaques, and interferes with clearance of neuritic plaques[52–55].

MS4A4A explains 3% of the local variance with negative correlation with AD risk (Fig 3c). Although the functional role of *MS4A* complex is largely unknown[52,53], recently *MS4A4A* was characterized as a novel surface marker for immature dendritic cells[56].

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

CaMMEL finds new genes in AD GWAS loci

We found 21 genes that explain high local genetic variance (at least 5%) in the GWAS / subthreshold regions (Fig 2a and Tab 1), of which 8 are located in the GWAS regions. A read-through gene *ATP5J2-PTCD1* mediates 12% of the local variance. The gene body is displaced from the strong GWAS peak in chromosome 7, but *cis*-eQTLs are located within the peak and share strong correlation within the LD (Fig 3a). Function of the readthrough protein is not well-understood, but *PTCD1* modulates mitochondrial precursor RNA and tRNA[57] and the related protein *PTCD2* was already recognized as a biomarker due to its elevated expressions in AD[58,59].

DPYSL2 is distal from the strong GWAS peak in chromosome 8, but its *cis*-eQTLs show a large degree of overlap with the GWAS region in LD (Fig 3b). *DPYSL2* was up-regulated in cortex, striatum and hippocampus after ischemic stroke[60], and it was listed among genes that modulate activation of microglial cells[61]. In fact, CaMMEL found that *DPYSL2* is positively correlated with AD progression.

ELMO1 is not co-localized with the strong GWAS peak in chromosome 7, yet *cis*-eQTLs are located within LD with GWAS SNPs (Fig 3d). This gene confers risk of immune disorders such as multiple sclerosis in previous GWAS[27,28], and triggers phagocytosis in microglia interacting with other proteins[26]. Our mediation analysis demonstrates that brain tissue-specific causal mechanism of *ELMO1* is negatively correlated with AD risk, and implicates mitigated activity of clearing neuritic plaques.

We note several cases where CaMMEL points to biologically important genes which are not captured by GWAS. For instance, *cis*-eQTLs of *CLP1* are located almost independent of the significant GWAS region on chromosome 11, but the gene's effect explains 11% of local variance (Fig 3c). *CLP1* is an interesting gene to follow up validation for neurodegenerative disorders; its mutation leads to damages in peripheral and central nervous systems altering tRNA synthesis[62,63], which conveys consistent implication as *ATP5J2-PTCD1*.

Two cytochrome P450 complex genes[64], *CYP27C1* (chr2) and *CYP39A1* (chr6) explain 15% and 7% of the local variance. *CYP27C1* is closely located with well-known *BIN1*, but the gene body does not overlap with strong GWAS signals. Instead, *cis*-eQTLs within 1Mb window cover that region and extend through non-zero LD effect of *CYP27C1* to nearly 2Mb around the GWAS locus (Fig 3e). The contribution of *BIN1* to AD is somewhat controversial[65–67], and we expect follow-up mediation analysis with cell type-specific eQTLs will further clarify the mechanisms.

The gene body of *CYP39A1* does not overlap with GWAS SNPs; however, *cis*-eQTLs are in LD with the GWAS variants (Fig 3f). *CYP27C1* is involved in lipid metabolism[68] and plays an important role in photoreceptors[69]. *CYP39A1* is involved in the generation of 25-hydroxy cholesterol from cholesterol. A regulatory role of hydroxy cholestrol in innate immunity is increasingly recognized by recent studies[70], and a murine experiment showed that inflammation in central nervous system due to disrupted lipid metabolism down-regulated activities of cytochrome P450 complex[71].

CaMMEL finds strong mediators in subthreshold loci

Of the 21 strong mediators, 13 are located in the subthreshold regions. We highlight three examples in the subthreshold regions. *SH3YL1* is located at the tip of chromosome 2 containing small number of genes and gene body is enclosed within subthreshold GWAS region (Fig 4a). *SH3YL1* was found significant in GWAS with attempted suicide and expressed in brain, but underlying mechanisms are unknown[72]. Interestingly this gene was also found robustly associated height in previous TWAS[17,18]. Several previous researches noted subtle relationship between AD and suicide attempt[73,74] and height[75]. However we think joint analysis with multiple GWAS traits will further edify detailed causal directions.

LRRC23 has *cis*-eQTLs in chromosome 12 between 7Mb and 8Mb that explain 7.8% of local genetic variance, followed by *CLSTN3* explains 1.6% (Fig 4b). Both mediation effects of *LRRC23* and *CLSTN3* are significantly non-zero. *LRRC23* and its *cis*-eQTLs were suggested as a modulator of innate immune network in mouse retina in

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

response to optic neuronal injury[24] and enriched in ramified microglial cells[76]. *CLSTN3* (calsyntenin-3) interacts with neurexin and forms synpatic adhesion[77] and up-regulated by increase of amyloid beta protein and accelerated neuronal death[78].

CNTFR explains 9.9% of genetic variance within 5Mb region (from 33Mb to 38Mb on chromosome 9) and multiple gene expression mediators are found (Fig 4c). Interestingly GWAS signals are also widely distributed with sporadic narrow peaks. *CNTFR*, interacting with *CNTF* and interferon gamma, stimulates murine microglia and increases expression of *CD40* on the cell surface[30]. Second best gene in the region, *EXOSC3* (RNA exsosome component 3), is also interesting. Exome sequencing followed by functional analysis in zebra fish identified that mutations in *EXOSC3* is causal to hypoplasia and spinal motor neuron degeneration[79], and expression of *EXOSC3* was differentially regulated in human monocyte and macrophage induced by lipopolysaccharide, suggesting functional roles in inflammatory bowel disease[80].

Discussion

We developed novel Bayesian machine learning approach to causal mediation analysis that infers regulatory mechanisms from summary statistics of GWAS and eQTLs. Our approach borrows great ideas from two existing statistical approaches, MR-Egger regression and TWAS. MR-Egger adjust for unmediated effects by meta-analysis of effect sizes[15], and TWAS correctly aggregate multiple SNP information taking into accounts of LD. However, these methods only provide partial solution of causal gene discoveries, posing strict assumptions such as independence between SNPs (MR-Egger) or giving up identifiability between mediated and unmediated effects (TWAS). On the contrary, CaMMEL addresses overall aspects of the causal mediation analysis in tight LD structure[36], by explicitly modeling both mediated and unmediated effects of multiple genes in a unified Bayesian framework and to identify causal pathways by carrying out Bayesian variable selection[38] between potentially collinear mediators[40,41]. Our direct Bayesian approach poses computational challenges on posterior probability calculation, but our efficient implementation made the method generally applicable to other related summary-based mediation and regression analysis.

Previous MR analysis on DNA methylation[81] used stringent FWER cutoff on instrument variable selection to avoid weak instrument bias observed in MR on individual-level data[48]. Here we retained *cis*-regulatory eQTLs and target genes with lenient p-value cutoff (0.05) for several reasons. Since CaMMEL estimates the RSS model and it reduces the risk of double counting evidences from multiple weak IVs. Forever, unlike endophenotypes in epidemiological studies such as Body Mass Index, strong genetic association of molecular QTL include effects from neighboring SNPs, and therefore stringent p-value cutoff does not make the QTL a strong IV[27,82,83]. In fact, our approach to analyze multiple mediators make causal mediators tested in more strict settings when we include as many competitor genes as possible.

We established top priority list of 21 genes as strong causal mediators of AD. Our results suggest contribution of innate immunity in AD progression pointing to a specific set of genes with proper brain tissue contexts. We expect our analysis framework can provide more complete pictures by combining with cell-type and tissue specific eQTL summary statistics. Currently we assume mediators (genes) act independently conditioned on genotype information, but extensions of LD blocks to pathways and multiple tissues would need more parsimonious models, for instance, by sharing information across tissues and genes of through factored regression models[84].

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Online methods

Review of regression with summary statistics

A summary association statistic refers to a univariate effect size (a regression slope of a simple regression or log-odds ratio in case-control studies) measured on each SNP without taking into account of LD structure. On a single trait GWAS, we normally have a vector of p summary statistics, effect size $\hat{\theta}_j$ and corresponding variance $\hat{\sigma}_j^2$ for each SNP $j \in [p]$[85]. However, due to LD (correlations between neighboring SNPs), an effect size measured on each single variant contains contributions from the neighboring SNPs. Unless we estimate causal variants by fine-mapping, inferring the multivariate effect from the univariate one, the univariate effect size of a SNP should not be interpreted as unbiased estimation of SNP-level signals[85,86].

Here our interest is on the multivariate (true) effect size vector $\boldsymbol{\theta}$ across all SNPs in the locus of interest. In principle, fixed effects can be modeled by a multivariate regression model, and parameters of this large regression model correspond to the multivariate effect $\boldsymbol{\theta}$. We assume an n-vector of individual-level phenotypes $\mathbf{y}$ was generated from $n \times p$ genotype matrix G on p SNPs with the multivariate effect sizes $\boldsymbol{\theta}$, and isotropic Gaussian noise parameter σ^2. More precisely,

$$\mathbf{y} \sim \mathcal{N}\big(G\boldsymbol{\theta}, \sigma^2 I\big)\,. \tag{4}$$

The regression with summary statistics (RSS) model[87,88] provides a principled way to describe a generative model of the GWAS summary data. We assume the observed GWAS summary effect sizes $\hat{\boldsymbol{\theta}} \equiv (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top$ were generated from the same multivariate effects $\boldsymbol{\theta}$ of individual-level multivariate GWAS model (eq. 4).

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\big(SRS^{-1}\boldsymbol{\theta}, SRS\big)\,, \tag{5}$$

where we denote the $p \times p$ linkage disequilibrium (LD) matrix R and the diagonal matrix of expected squares, S with each element $S_{jj} = \sqrt{\mathbb{E}[\hat{\theta}_j^2]} = \sqrt{\hat{\theta}_j^2/n + \sigma_j^2}$. In practice we estimate the LD matrix $\hat{R}$ from a standardized genotype matrix of reference cohort by taking crossproduct, $\hat{R} = G^\top G/(n-1)$. While other methods regularize $\hat{R}$ matrix by adding ridge regression penalty on the diagonal matrix[89], here we regularized $\hat{R}$ by dropping variance components corresponding to small eigenvalues (λ)[90]. We use a fixed cutoff $\lambda < 10^{-2}$, but our results were robust to the different cutoff values such as $\lambda < 10^{-3}$ and $\lambda < 10^{-1}$.

Derivation of the causal mediation model in RSS

We derive the mediation model (eq. 2), assuming that the observed univariate eQTL effects on a mediator k, $\hat{\boldsymbol{\alpha}}_k \equiv (\hat{\alpha}_{1k}, \ldots, \hat{\alpha}_{pk})^\top$ are generated from the RSS model with mean $\mathbb{E}[\hat{\boldsymbol{\alpha}}_k] = S_\alpha R S_\alpha^{-1}\boldsymbol{\alpha}_k$ with expected squares S_α of eQTL effects. Here we assume standard errors of GWAS effects can be substituted with standard errors of QTL effects up to some scaling factor, i.e., $S_\alpha = cS$ with some real number $c > 0$, because association statistic is mainly determined by the underlying minor allele frequencies and samples sizes, and this allows us to rewrite the expectation of mediation effect (eq. 1):

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \left(S_\alpha R S_\alpha^{-1}\boldsymbol{\alpha}_k\right)\beta_k + SRS^{-1}\boldsymbol{\gamma}$$

where the constant factor c cancels. With no measurement error, i.e., $\mathbb{E}[\hat{\boldsymbol{\alpha}}] = \hat{\boldsymbol{\alpha}}$, we can reasonably assume:

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] \approx \sum_k \hat{\boldsymbol{\alpha}}_k \beta_k + SRS^{-1}\boldsymbol{\gamma}.$$

Substitution of this mean effect to the RSS model (eq. 5) yields the probabilistic model for CaMMEL (eq. 2).

Variational Bayes inference of the mediation model

A key challenge in fitting the causal mediation model (eq. 2) is dealing with the covariance matrix in the likelihood. We exploit the spectral transformation of the LD matrix to make the model inference more amenable[91]. We take singular value decomposition of standardized genotype matrix, $(n)^{-1/2}G = UDV^\top$, and decompose LD matrix into $\hat{R} = VD^2V^\top$, exposing that the effective number of sample size $\tilde{n}$ is bounded by the sample size of reference panel, $\tilde{n} < n$, after the regularization $\hat{R}$. Defining $\tilde{\mathbf{y}} \equiv V^\top S^{-1}\hat{\boldsymbol{\theta}}$ we obtain equivalent, but fully factorized, multivariate Gaussian model

$$\tilde{\mathbf{y}} \sim \mathcal{N}\left(\sum_{k=1}^{K} V^\top S^{-1}\hat{\boldsymbol{\alpha}}_k\beta_k + D^2V^\top S^{-1}\boldsymbol{\gamma}, D^2\right). \tag{6}$$

We efficiently fit the model using black box variational inference[92,93] with a novel reparameterization trick[84,94]. Normally a high-dimensional multivariate regression is intractable problem since the total amount of model variance blows up as a function of dimensions (p). However, instead of dealing with large number of parameters (p SNPs), we deal with smaller $\tilde{n}$-dimensional aggregate random variables ($\tilde{n} < n \ll p$), a linear combination of p-dimensional effects. More precisely we define

$$\eta_i \equiv \sum_{j=1}^{p} V_{ji}S_j^{-1}\gamma_j, \quad \xi_i \equiv \sum_{j=1}^{p} V_{ji}S_j^{-1}\sum_{k=1}^{K}\hat{\alpha}_{jk}\beta_k$$

to rewrite the transformed log-likelihood of the model for each eigen component i:

$$\ln P(\tilde{y}_i|\eta_i,\xi_i) = -\frac{1}{2}\ln d_i^2 - \frac{1}{2d_i^2}(\tilde{y}_i - d_i^2\eta_i - \xi_i)^2 - \frac{1}{2}\ln(2\pi).$$

This reformulation not only achieves faster convergence by reducing variance[94,95], but we also gain computational efficiency that we can sample all the eigen components independently in parallel taking full accounts of the underlying LD structure between SNPs.

The overall algorithm proceeds as follows: We first update surrogate distributions of $q(\xi) \approx \mathcal{N}(\mu_\xi, \nu_\xi)$ and $q(\eta) \approx \mathcal{N}(\mu_\eta, \nu_\eta)$ by minimizing Kullback-Leibler (KL) divergence between the surrogate q and true distribution P, $D_{\mathrm{KL}}(q\|P)$, by taking stochastic gradient steps with respect to the mean $\nabla_{\mu_\xi}, \nabla_{\mu_\eta}$ and variance $\nabla_{\nu_\xi}, \nabla_{\nu_\eta}$[93]; we then then back-propagate this gradient to the gradient with respect to the original mean $\nabla_{\mu_\beta}, \nabla_{\mu_\gamma}$ and variance parameters $\nabla_{\nu_\beta}, \nabla_{\nu_\gamma}$ to eventually find $q(\beta) \approx \mathcal{N}(\mu_\beta, \nu_\beta)$ and $q(\gamma) \approx \mathcal{N}(\mu_\gamma, \nu_\gamma)$. We formulate the variational mean μ_β and variance ν_β of the spike-slab distribution following the previous derivations[39] (see our technical paper[84] for details).

Calculation of proportion of variance explained

We measured explained variance according to the definition of Shi *et al.*[90] Total variance can be decomposed into mediated component, $(\boldsymbol{\alpha\beta})^\top R(\boldsymbol{\alpha\beta})$ and unmediated component $(\boldsymbol{\gamma})^\top R(\boldsymbol{\gamma})$. We checked this provides a reasonably tight lower-bound of actual model variance through simulations. Parameters, α, β, γ, are estimated by posterior mean of variational Bayes inference.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Parametric bootstrap

Variational inference may only capture one mode of the true posterior, and so tends to underestimate posterior variances. Therefore, we do not rely on the Bayesian posterior variance to assess confidence in the estimated mediated and unmediated effect sizes. Moreover, the effects tend to correlate with each other then can lead to a severe collinearity problem. Instead, we use the parametric bootstrap to estimate the null distribution of these parameters and calibrate false discovery rates[43,44,46].

Confounder correction in gene expression matrix

We learn and adjust for unobserved non-genetic confounders in gene expression by half-sibling regression (Supp Fig 2)[35]. For each gene we find 105 control genes, selecting the most correlated 5 genes from each other chromosome. We assume *cis*-regulatory effects only occur within each chromosome and that inter-chromosomal correlations are due only to shared non-genetic confounders. This assumption is invalid when *cis*-regulatory variants under consideration have *trans*-effects on the control genes. Therefore, we regress out genetic effects on the established control genes. Half-sibling regression simply takes residuals after regressing out the control gene effects on observed gene expressions.

Establishing the extended LD blocks

We establish independent LD blocks of SNPs within a chromosome and restrict mediation analysis within each block. We computed pairwise correlations between pairs of variants in the Thousand Genomes European samples within 1 megabase and with $r^2 > 0.1$. We pruned to a desired threshold by iteratively picking the top-scoring AD GWAS variant (breaking ties arbitrarily) and removing the tagged variants until no variants remained. At convergence each small block is tagged by its top-scoring variant, and merge LD blocks overlapping in genome.

Simulation framework

We used 66 real extended loci, randomly sampling 3 per chromosome that each harbored more than 1,000 SNPs. Within each region, we assigned each non-overlapping window of 20 SNPs to a synthetic gene. For each gene, we varied the number of causal n_q SNPs (per gene) and generated n_g gene expression data from a linear model with proportion of variance mediated through expression (PVE) from $\tau = 0.01$ to 0.2. For each region, we selected n_d SNPs to have a direct (unmediated) effect on the phenotype and selected genes to have a causal effect on the phenotype, varying the variance explained by the direct effect δ (horizontal pleiotropy[13]). For simplicity we used $n_d = n_q$, meaning direct effect contributes variance of one gene, and gene expression heritability was fixed to $\rho = 0.17$ as used in our previous work[96]. We carefully selected pleiotropic SNPs from the eQTLs of non-mediating genes. We evaluated the statistical power of each method at fixed FDR 1% and the precision by the area under the precision recall curve (AUPRC).

We used a generative scheme proposed by Mancuso *et al.*[18] with a slight modification that includes unmediated genetic effects on the trait. Let n_q be number of causal SNPs per each genetic gene; n_g be number of causal mediation genes; n_d be number of SNPs driving unmediated effect on trait. For each simulation, three PVE (a proportion of variance explained) parameters are given between 0 and 1. (1) ρ: PVE of gene expression; (2) τ: PVE of mediation; (3) δ: PVE of unmediated effect.

For each mediator k on individual i: $M_{ik} = \sum_{j=1}^{n_q} G_{i(j)}\alpha_{jk} + \epsilon_{ik}$ where (j) is j-th causal SNP index for this gene, $\alpha \sim \mathcal{N}(0, \rho/n_q)$ and $\epsilon \sim \mathcal{N}(0, 1-\rho)$.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Combining stochastically generated mediator variables, we generate phenotypes

$$y_i = \sum_{k=1}^{n_g} M_{i(k)}\beta_k + \sum_{j'=1}^{n_d} G_{i(j')}\gamma_{j'} + \epsilon_i$$

where (k) is k-th causal mediator index, mediation effect sizes are sampled independently $\beta \sim \mathcal{N}(0, \tau/n_g)$, and unmediated effect sizes are also sampled independently $\gamma \sim \mathcal{N}(0, \delta/n_d)$ and we infused irreducible noise to take into account of the residual variance $\epsilon \sim \mathcal{N}(0, 1-\tau-\delta)$.

Calculation of TWAS test statistics from summary z-scores

For simplicity we used completely summary-based approach[18] that need not pretrain multivariate regression model. TWAS statistics can be derived from two sets of summary z-score vectors–one for GWAS $\mathbf{z}_G$ and the other for eQTL $\mathbf{z}_T$ using the fact that multivariate QTL effect size vector α can be approximated by LD-adjusted effect, $\boldsymbol{\alpha} \approx R^{-1}\mathbf{z}_T/\sqrt{n}$.

$$T = \frac{\mathbf{z}_G^\top \boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^\top R \boldsymbol{\alpha} + \lambda}} \tag{7}$$

Test statistics T asymptotically follow $\mathcal{N}(0,1)$. We set $\lambda = 10^{-8}$ to avoid zero divided by zero.

Data preprocessing

We used genotypes of 672,266 SNPs in 1,709 individuals from the Religious Orders Study (ROS) and the Memory and Aging Project (MAP)[97]. We mapped hg18 coordinates of SNPs (Affymetrix GeneChip 6.0) to hg19 coordinates matching strands using publicly available information (http://www.well.ox.ac.uk/~wrayner/strand/GenomeWideSNP_6.na32-b37.strand.zip). We imputed the genotype arrays by prephasing haplotypes based on the 1000 genome project phase I version 3[98] using `SHAPEIT`[99], retaining only SNPs with MAF > 0.05. We then imputed SNPs in 5MB windows using `IMPUTE2`[100] with 100 Markov Chain Monte Carlo iterations and 10 burn-in iterations. After the full imputation, 6,516,083 SNPs were considered in follow-up expression and methylation QTL analysis.

We used gene expression data generated by RNA-seq from dorsolateral prefrontal cortex (DLPFC) of 540 individuals[4]. We retained 436 samples by first removing potentially poor quality 84 samples with RIN score below 6 (suggested by the GTEx consortium) and further removing 20 samples with no amyloid beta measurements. Of 436 samples, we used 356 samples with genotype information for eQTL analysis. Gene-level quantification was conducted by `RSEM`[101] and we focused on 18,461 coding genes out of 55,889 according to the GENCODE annotations (v19 used in the GTEx v6p[102]).

Original gene-level quantification data follows negative binomial distribution. We adjusted variability of sequencing depth across samples following previous methods[103,104]. We then converted over-dispersed count data to Normal distribution data while fitting gene by gene null model of negative binomial regression that only includes intercept term, equivalent to `rlog` transformation[105]. We used custom-designed inference algorithm to speed up the inference of models (https://github.com/ypark/fqtl). We found residual values inside the inverse link function follow approximately Normal distribution in most genes.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

GWAS summary statistics

International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tables

Table 1. CaMMEL identified 21 genes explains at least 5% of local variance in subthreshold / GWAS regions. All genes in the list passed stringent p-value threshold < 3e-06 (FDR < .007%) and GWAS regions contain at least one subthreshold or more significant SNP (p-value < 10^{-4}). *Gene*: gene symbol in `hg19`; *chr*: chromosome name; *TSS*: transcription start site with respect to strand (kb); *TES*: transcription end site (kb); *Mediation*: gene expression mediation effect size with standard deviation and posterior inclusion probability (first and second numbers in the bracket). *Best GWAS*: best GWAS ID in the LD block with z-score and location in kb. *Best QTL*: best eQTL SNP ID with z-score and location in kb. *PVE*: proportion of local genetic variance explained by gene expression mediation (in %).

Gene	chr	TSS	TES	Mediation	Best GWAS	Best QTL	PVE
SH3YL1	2	266	218	0.22 (0.016, 1)	rs75141812 (-4, 290)	rs1474053 (-5.4, 224)	8.3
CYP27C1	2	127,978	127,942	0.33 (0.017, 1)	rs6733839 (14, 127,893)	rs6430934 (-3.8, 128,001)	15
CYP39A1	6	46,621	46,518	-0.27 (0.032, 1)	rs10948363 (6.6, 47,488)	rs9296505 (-4.1, 46,630)	7
RGS17	6	153,452	153,326	0.21 (0.035, 1)	rs9479690 (4.2, 154,072)	rs12526771 (-3, 153,783)	7.4
ELMO1	7	37,489	36,894	-0.28 (0.016, 1)	rs2718058 (-5.9, 37,842)	rs80195177 (2.8, 37,637)	7.6
ATP5J2-PTCD1	7	99,039	99,039	0.23 (0.012, 1)	rs12539172 (-6.2, 100,092)	rs2525546 (-3.3, 99,552)	12
COPS6	7	99,687	99,690	-0.21 (0.012, 1)	rs12539172 (-6.2, 100,092)	rs4308665 (-3.5, 99,667)	8.9
DPYSL2	8	26,372	26,516	0.39 (0.016, 1)	rs7982 (-10, 27,462)	rs6558007 (-2.4, 27,434)	7.7
CLU	8	27,473	27,454	0.18 (0.014, 1)	rs7982 (-10, 27,462)	rs11136000 (3.7, 27,465)	5.9
FSBP	8	95,449	95,449	0.26 (0.044, 1)	rs7818382 (5.4, 96,054)	rs79027703 (-2.9, 95,664)	7.5
C8orf37	8	96,281	96,257	0.16 (0.017, 1)	rs7818382 (5.4, 96,054)	rs2514558 (-5, 96,292)	5.4
CNTFR	9	34,590	34,551	0.14 (0.039, 1)	rs7040732 (4.1, 38,488)	rs149113257 (3.4, 33,670)	9.9
RHOBTB1	10	62,761	62,629	0.14 (0.019, 1)	rs10761556 (-4, 62,518)	rs1372711 (3.3, 62,431)	7.2
CLP1	11	57,424	57,429	0.16 (0.032, 1)	rs983392 (-8.1, 59,924)	rs7102963 (3.1, 57,027)	11
LRRC23	12	6,993	7,023	-0.096 (0.01, 1)	rs11064497 (-4.5, 7,170)	rs12315375 (8.1, 7,009)	7.8
FRS2	12	69,864	69,974	-0.091 (0.015, 1)	rs4351896 (4.3, 69,353)	rs2601007 (-3.5, 69,979)	9.7
PTPN21	14	89,021	88,932	-0.16 (0.016, 1)	rs12433739 (4.2, 88,798)	rs59261166 (2.6, 88,814)	6.7
EMC4	15	34,517	34,522	0.1 (0.018, 1)	rs112929571 (3.9, 34,424)	rs59209438 (-9.1, 34,536)	6.8
LEO1	15	52,264	52,230	0.13 (0.013, 1)	rs8035452 (-5.1, 51,041)	rs6493549 (-4.3, 52,528)	8
DENND4A	15	66,085	65,950	0.17 (0.02, 1)	rs74615166 (5.1, 64,725)	rs7178388 (-2.5, 66,957)	5.1
KLK5	19	51,456	51,447	-0.18 (0.023, 1)	rs12459419 (-5.4, 51,728)	rs7253829 (2.3, 51,754)	6.3

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

Figure 1. Schematic of the CaMMEL method. **(a)** Transcriptome-wide causal mediation analysis overcomes limitation of observed TWAS by taking advantages of large statistical power of Alzheimer's disease GWAS statistics and brain tissue-specific regulatory contexts of eQTL data. **(b)** Correlation between gene expression and disease status can arise from mediation, pleiotropy and reverse causation. **(c)** CaMMEL jointly estimates two regression models taking into account multiple sources of gene expression variation.

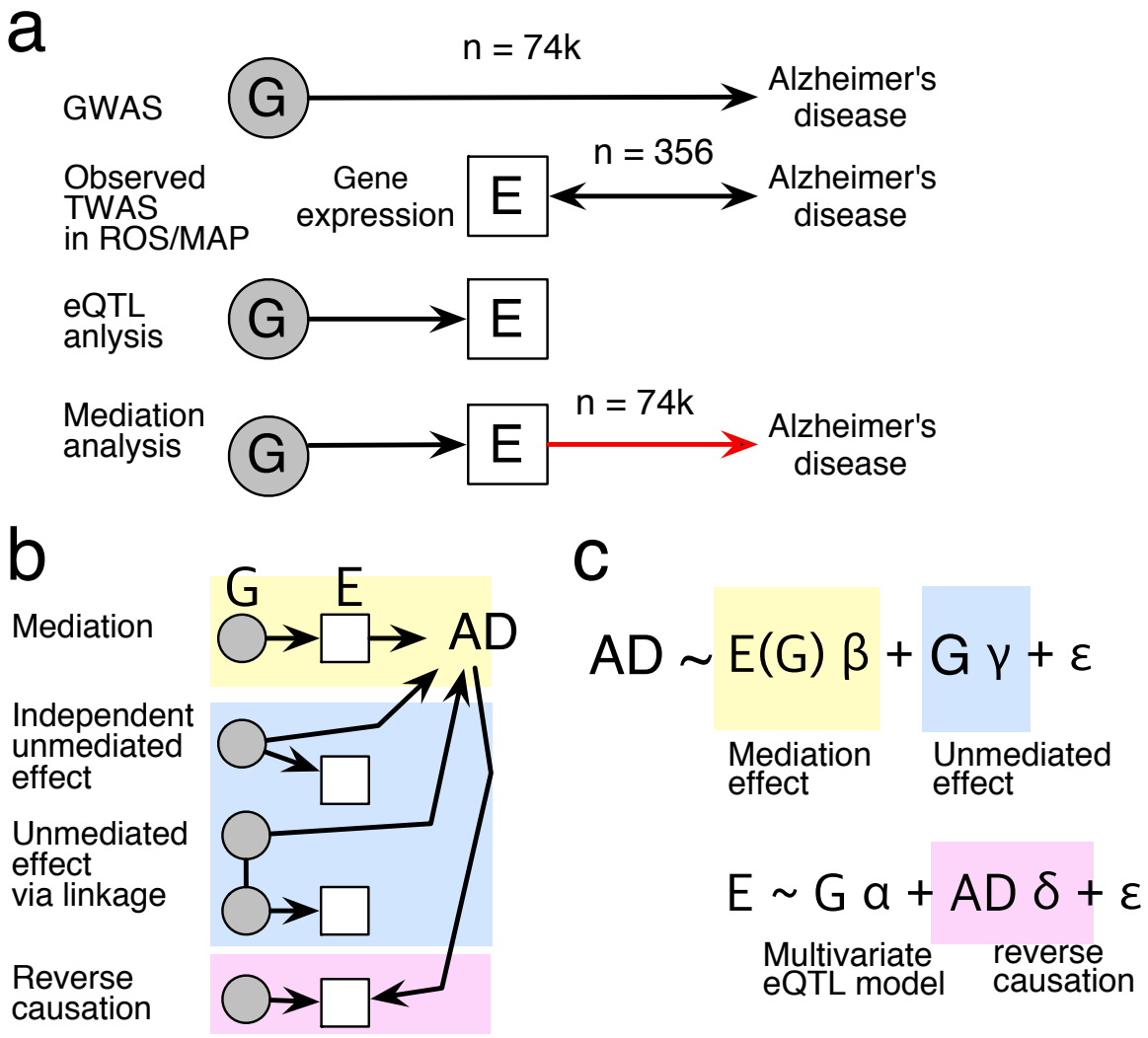


Figure 1: Overview of study design and methods

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2. Transcriptome-wide mediation analysis reveals causal target genes in AD. **(a)** Mediation effect sizes over all the genes located in GWAS subthreshold region ($p < 10^{-4}$). Size of dots are proportional to the proportion of local genetic variance explained by mediation. Errorbars show 95% confidence intervals. Dots are colored according to the GWAS significance. 21 genes (PVE > 5%) are annotated (Table 1). **(b)** Histogram showing proportion of variance explained (PVE) by gene expression mediation. *Gray bars*: total histogram of PVE; *green bars*: histogram of PVE for 774 significant mediation genes (non-zero mediation FDR < 10^{-4}); *blue line*: histogram of PVE for 206 significant mediation genes found in GWAS subthreshold regions (FDR < 10^{-4}, GWAS p < 10^{-4}). *Green horizontal line*: average PVE for 774 genes. *Red horizontal line*: 5% cutoff for strong mediators. **(c)** Density estimation showing relationship between PVE (y-axis) and GWAS significance level (x-axis). The 21 strong mediators are shown as red dots. *Red horizontal line*: 5% cutoff. **(d)** Density estimation showing relationship between posterior probability of mediation (y-axis) and GWAS significance level (x-axis). We marked 21 strong genes with red dots. *Red horizontal line*: posterior probability corresponding to FDR < 10^{-4} cutoff. **(e)** Density estimation showing relationship between best eQTL (y-axis) and best GWAS significance level (x-axis) within LD.

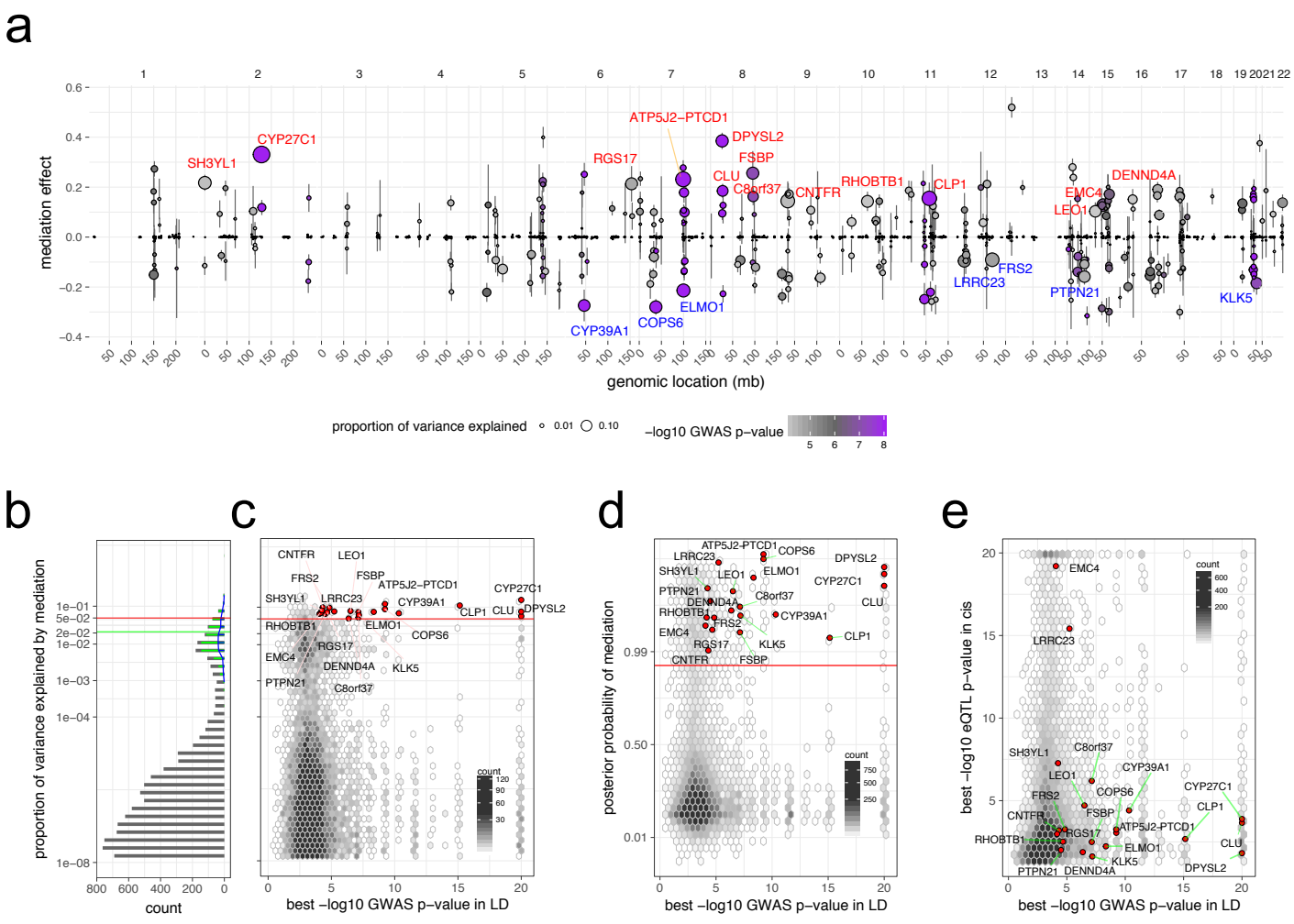


Figure 2: Genome-wide patterns of gene expression mediation in AD.

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3. Local views of significant gene expression mediation on GWAS significant regions. In each plot, proportion of gene expression mediation (%) and GWAS Manhattan plot (upside down) are aligned according to genomic locations. eQTL links are colored based on z-score of effect sizes (blue = negative; red = positive). Ranges of correlation due to LD were covered with green shades. Gene bodies are marked by dark green bars in genomic location. Numbers within brackets denote mediation test p-value.

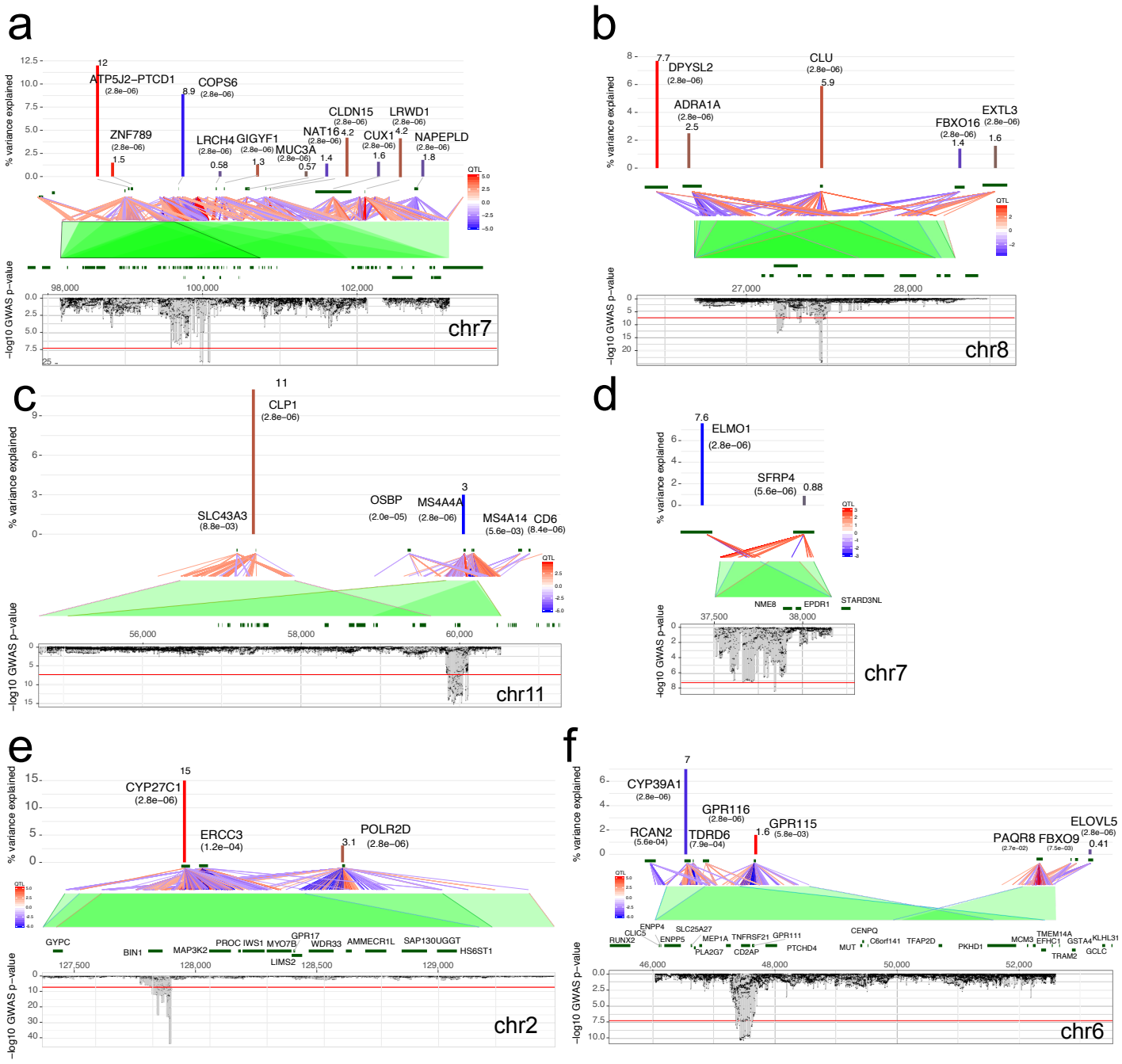


Figure 3: Examples in the GWAS regions

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4. Local views of significant gene expression mediation on subthreshold regions. In each plot, proportion of gene expression mediation (%) and GWAS Manhattan plot (upside down) are aligned according to genomic locations. eQTL links are colored based on z-score of effect sizes (blue = negative; red = positive). Ranges of correlation due to LD were covered with green shades. Gene bodies are marked by dark green bars in genomic location. Numbers within brackets denote mediation test p-value.

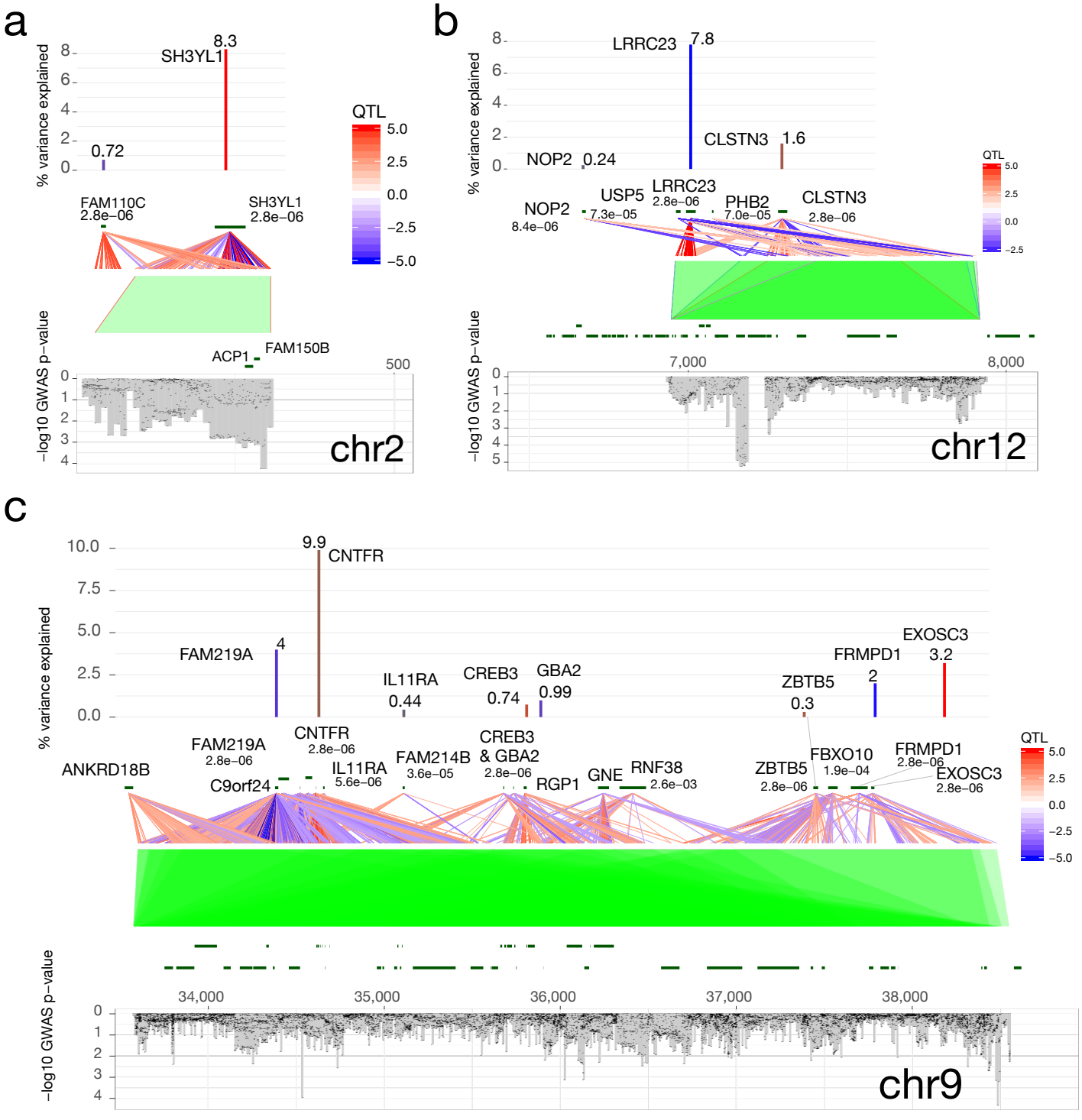


Figure 4: Examples in the sub-threshold regions

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgment

We thank Alkes Price (Harvard), Alexander Gusev (Dana Farber) and Bogdan Pasaniuc (UCLA) for insightful discussion.

We have source code for CaMMEL made publicly available as a part of the `R` package for summary-based QTL / GWAS analysis (`https://github.com/YPARK/zqtl`). Effect sizes of CaMMEL are estimated through `fit.med.zqtl` function.

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i–Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Universite de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant 503480), Alzheimer's Research UK (Grant 503176), the Wellcome Trust (Grant 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728.

Author contribution

MK and PLD conceived study design. PLD provided ROS/MAP gene expression and DNA methylation data. YP, AKS and LH developed the method. YP implemented `C++` and `R` software. YP carried out mediation analysis. JD helped post-processing. YP, AKS, LH, JD, PLD and MK wrote the manuscript.

References

1. Hardy, J. & Selkoe, D. J. The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science (New York, N.Y.)* **297,** 353–356 (2002).

2. Musiek, E. S. & Holtzman, D. M. Three dimensions of the amyloid hypothesis: time, space and 'wingmen'. *Nature Neuroscience* **18,** 800–806 (2015).

3. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153,** 707–720 (2013).

4. Mostafavi, S. *et al.* A molecular network of the aging brain implicates INPPL1 and PLXNB1 in Alzheimer's disease. *bioRxiv* 205807 (2017). doi:10.1101/205807

5. Raj, T. *et al.* Integrative analyses of splicing in the aging brain: role in susceptibility to Alzheimer's Disease. *bioRxiv* 174565 (2017). doi:10.1101/174565

6. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics* **45,** 1452–1458 (2013).

7. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. *The*

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

American Journal of Human Genetics **93,** 779–797 (2013).

8. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine* **373,** 895–907 (2015).

9. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538,** 523–527 (2016).

10. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports* **17,** 2042–2059 (2016).

11. Smith, G. D. & Ebrahim, S. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33,** 30–42 (2004).

12. Katan, M. B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *International Journal of Epidemiology* **33,** 9–9 (2004).

13. Smith, G. D. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics* **23,** ddu328–R98 (2014).

14. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14,** 483–495 (2013).

15. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512–525 (2015).

16. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47,** 1091–1098 (2015).

17. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48,** 245–252 (2016).

18. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *The American Journal of Human Genetics* **100,** 473–487 (2017).

19. Wainberg, M. *et al.* Vulnerabilities of transcriptome-wide association studies. *bioRxiv* 206961 (2017). doi:10.1101/206961

20. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature Neuroscience* **511,** 421 (2017).

21. Pelham, C. J. *et al.* Cullin-3 Regulates Vascular Smooth Muscle Function and Arterial Blood Pressure via PPARγ and RhoA/Rho-Kinase. *Cell Metabolism* **16,** 462–472 (2012).

22. Basson, M. Cardiovascular diseases: Degradation relieves the pressure. *Nature Medicine* **18,** nm.3007–1629 (2012).

23. Sweeney, G. & Song, J. The association between PGC-1α and Alzheimer's disease. *Anatomy & Cell Biology* **49,** 1–6 (2016).

24. Templeton, J. P. *et al.* Innate Immune Network in the Retina Activated by Optic Nerve Crush. *Investigative Ophthalmology & Visual Science* **54,** 2599–2606 (2013).

25. Nevers, Y. *et al.* Insights into Ciliary Genes and Evolution from Multi-Level Phylogenetic Profiling. *Molecular Biology and Evolution* **34,** 2016–2034 (2017).

26. Sierra, A., Abiega, O., Shahraz, A. & Neumann, H. Janus-faced microglia: beneficial and detrimental consequences of microglial phagocytosis. *Frontiers in Cellular Neuroscience* **7,** (2013).

27. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature genetics* **advance online publication SP - EP -,** (2017).

28. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics* **45,** 1353–1360 (2013).

29. Hausmann, O. N. *et al.* Spinal cord injury induces expression of RGS7 in microglia/macrophages in rats. *European Journal of Neuroscience* **15,** 602–612 (2002).

30. Lin, H.-W., Jain, M. R., Li, H. & Levison, S. W. Ciliary neurotrophic factor (CNTF) plus soluble CNTF receptor α increases cyclooxygenase-2 expression, PGE 2 release and interferon-γ-induced CD40 in murine microglia. *Journal of Neuroinflammation* **6,** 7 (2009).

31. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimers disease. *Nature* **518,** 365–369

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2015).

32. Pearl, J. Interpretation and identification of causal mediation. *Psychological methods* **19,** 459–481 (2014).

33. VanderWeele, T. & Vansteelandt, S. Mediation Analysis with Multiple Mediators. *Epidemiologic Methods* **2,** 1–22 (2013).

34. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic epidemiology* **37,** 658–665 (2013).

35. SchÃűlkopf, B. *et al.* Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences* **113,** 7391–7398 (2016).

36. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48,** 481–487 (2016).

37. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics* **9,** e1003264 (2013).

38. Mitchell, T. J. & Beauchamp, J. J. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* **83,** 1023–1032 (1988).

39. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis* **7,** 73–108 (2012).

40. RoÄ■kovÃą, V. & George, E. I. Negotiating multicollinearity with spike-and-slab priors. *METRON* **72,** 217–229 (2014).

41. Ghosh, J. & Ghattas, A. E. Bayesian Variable Selection Under Collinearity. *The American Statistician* **69,** 165–173 (2015).

42. Wasserman, L. All of Statistics: A Concise Course in Statistical Inference. 442 (2010). at <http://dl.acm.org/citation.cfm?id=1965575>

43. Freedman, D. & Lane, D. A Nonstochastic Interpretation of Reported Significance Levels. *Journal of Business & Economic Statistics* **1,** 292 (1983).

44. BůÅ¿kovÃą, P., Lumley, T. & Rice, K. Permutation and Parametric Bootstrap Tests for GeneGene and GeneEnvironment Interactions. *Annals of Human Genetics* **75,** 36–45 (2011).

45. Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *NeuroImage* **92,** 381–397 (2014).

46. Wahl, S. *et al.* On the potential of models for location and scale for genome-wide DNA methylation data. *BMC bioinformatics* **15,** 1 (2014).

47. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100,** 9440–9445 (2003).

48. Burgess, S. & Thompson, S. G. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology* **40,** 755–764 (2011).

49. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).

50. Liu, Y., Schiff, M., Serino, G., Deng, X.-W. & Dinesh-Kumar, S. P. Role of SCF Ubiquitin-Ligase and the COP9 Signalosome in the N GeneMediated Resistance Response to Tobacco mosaic virus. *The Plant Cell* **14,** 1483–1496 (2002).

51. Deng, Z. *et al.* Plant homologue constitutive photomorphogenesis 9 (COP9) signalosome subunit CSN5 regulates innate immune responses in macrophages. *Blood* **117,** 4796–4804 (2011).

52. Karch, C. M., Cruchaga, C. & Goate, A. M. Alzheimers Disease Genetics: From the Bench to the Clinic. *Neuron* **83,** 11–26 (2014).

53. Villegas-Llerena, C., Phillips, A., Garcia-Reitboeck, P., Hardy, J. & Pocock, J. M. Microglial genes regulating neuroinflammation in the progression of Alzheimer's disease. *Current Opinion in Neurobiology* **36,** 74–81 (2016).

54. Nuutinen, T., Suuronen, T., Kauppinen, A. & Salminen, A. Clusterin: A forgotten player in Alzheimer's disease. *Brain Research Reviews* **61,** 89–104 (2009).

55. Li, X. *et al.* Clusterin in Alzheimers disease: a player in the biological behavior of amyloid-beta. *Neuroscience Bulletin* **30,** 162–168 (2013).

56. Mansoor, A. *et al.* MS4A4A: a novel cell surface marker for M2 macrophages and plasma cells. *Immunology and Cell Biology* **95,**

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

611–619 (2017).

57. Rackham, O. *et al.* Pentatricopeptide repeat domain protein 1 lowers the levels of mitochondrial leucine tRNAs in cells. *Nucleic Acids Research* **37,** 5859–5867 (2009).

58. Lightowlers, R. N. & Chrzanowska-Lightowlers, Z. M. Human pentatricopeptide proteins. *RNA Biology* **10,** 1433–1438 (2013).

59. Nagele, E., Han, M., DeMarshall, C., Belinka, B. & Nagele, R. Diagnosis of Alzheimer's Disease Based on Disease-Specific Autoantibody Profiles in Human Sera. *PloS one* **6,** e23112 (2011).

60. Indraswari, F., Wong, P. T. H., Yap, E., Ng, Y. K. & Dheen, S. T. Upregulation of Dpysl2 and Spna2 gene expression in the rat brain after ischemic stroke. *Neurochemistry International* **55,** 235–242 (2009).

61. Juknat, A. *et al.* Microarray and Pathway Analysis Reveal Distinct Mechanisms Underlying Cannabinoid-Mediated Modulation of LPS-Induced Activation of BV-2 Microglial Cells. *PloS one* **8,** e61462 (2013).

62. Hanada, T. *et al.* CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature* **495,** 474–480 (2013).

63. Karaca, E. *et al.* Human CLP1 Mutations Alter tRNA Biogenesis, Affecting Both Peripheral and Central Nervous System Function. *Cell* **157,** 636–650 (2014).

64. Nebert, D. W., Wikvall, K. & Miller, W. L. Human cytochromes P450 in health and disease. *Philos Trans R Soc Lond B Biol Sci* **368,** 20120431–20120431 (2013).

65. De Rossi, P. *et al.* BIN1 localization is distinct from Tau tangles in Alzheimers disease. *Matters* **3,** e201611000018 (2017).

66. Adams, S. L., Tilton, K., Kozubek, J. A., Seshadri, S. & Delalle, I. Subcellular Changes in Bridging Integrator 1 Protein Expression in the Cerebral Cortex During the Progression of Alzheimer Disease Pathology. *Journal of Neuropathology & Experimental Neurology* **75,** 779–790 (2016).

67. Calafate, S., Flavin, W., Verstreken, P. & Moechars, D. Loss of Bin1 Promotes the Propagation of Tau Pathology. *Cell Reports* **17,** 931–940 (2017).

68. Lathe, R., Sapronova, A. & Kotelevtsev, Y. Atherosclerosis and Alzheimer - diseases with a common cause? Inflammation, oxysterols, vasculature. *BMC Geriatrics* **14,** 7 (2014).

69. Enright, J. M. *et al.* Cyp27c1 Red-Shifts the Spectral Sensitivity of Photoreceptors by Converting Vitamin A1 into A2. *Current Biology* **25,** 3048–3057 (2015).

70. Cyster, J. G., Dang, E. V., Reboldi, A. & Yi, T. 25-Hydroxycholesterols in innate and adaptive immunity. *Nature Reviews Immunology* **14,** nri3755–743 (2014).

71. Renton, K. W. & Nicholson, T. E. Hepatic and Central Nervous System Cytochrome P450 Are Down-Regulated during Lipopolysaccharide-Evoked Localized Inflammation in Brain. *Journal of Pharmacology and Experimental Therapeutics* **294,** 524–530 (2000).

72. Willour, V. L. *et al.* A genome-wide association study of attempted suicide. *Molecular psychiatry* **17,** 433–444 (2011).

73. Barak, Y. & Aizenberg, D. Suicide amongst Alzheimers Disease Patients: A 10-Year Survey. *Dementia and Geriatric Cognitive Disorders* **14,** 101–103 (2002).

74. Seyfried, L. S., Kales, H. C., Ignacio, R. V., Conwell, Y. & Valenstein, M. Predictors of suicide in patients with dementia. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* **7,** 567–573 (2017).

75. Schnaider Beeri, M. Relationship Between Body Height and Dementia. *American Journal of Geriatric Psychiatry* **13,** 116–123 (2005).

76. Parakalan, R. *et al.* Transcriptome analysis of amoeboid and ramified microglia isolated from the corpus callosum of rat brain. *BMC neuroscience* **13,** 64 (2012).

77. Lu, Z. *et al.* Calsyntenin-3 molecular architecture and interaction with neurexin 1α. *Journal of Biological Chemistry* **289,** 34530–34542 (2014).

78. Uchida, Y., Nakano, S.-i., Gomi, F. & Takahashi, H. Up-regulation of calsyntenin-3 by β-amyloid increases vulnerability of cortical neurons. *FEBS Letters* **585,** 651–656 (2011).

79. Wan, J. *et al.* Mutations in the RNA exosome component gene EXOSC3 cause pontocerebellar hypoplasia and spinal motor neuron degeneration. *Nature genetics* **44,** 704 EP ––708 (2012).

80. Baillie, J. K. *et al.* Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

insights into genetic aetiology of inflammatory bowel disease. *PLoS genetics* **13,** e1006641 (2017).

81. Richardson, T. G. *et al.* Causal epigenome-wide association study identifies CpG sites that influence cardiovascular disease risk. *bioRxiv* 132019 (2017). doi:10.1101/132019

82. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS genetics* **10,** e1004383 EP (2014).

83. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *bioRxiv* 065037 (2016). doi:10.1101/065037

84. Park, Y., Sarkar, A. K., Bhutani, K. & Kellis, M. Multi-tissue polygenic models for transcriptome-wide association studies. *bioRxiv* 107623 (2017). doi:10.1101/107623

85. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews Genetics* **18,** 117–127 (2016).

86. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30,** 2906–2914 (2014).

87. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *bioRxiv* 042457 (2016). doi:10.1101/042457

88. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association. *Genetics* **198,** 497–508 (2014).

89. Wen, X. & Stephens, M. Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to geneenvironment interactions. *The Annals of Applied Statistics* **8,** 176–203 (2014).

90. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99,** 139–153 (2016).

91. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature methods* **8,** 833–835 (2011).

92. Ranganath, R., Gerrish, S. & Blei, D. M. Black Box Variational Inference. in *Proceedings of the 13th international conference on artificial intelligence and statistics* (eds. Kaski, S. & Corander, J.) 814–822 (2014). at <http://jmlr.org/proceedings/papers/v33/ranganath14.pdf>

93. Paisley, J. W., Blei, D. M. & Jordan, M. I. Stick-breaking beta processes and the Poisson process. in *Proceedings of the 14th international conference on artificial intelligence and statistics* (2012). at <http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2012_PaisleyBJ12.pdf>

94. Wang, S. & Manning, C. Fast dropout training. *Proceedings of the 30th International Conference on Machine Learning* **28,** 118–126 (2013).

95. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick. *arXiv.org* (2015). at <http://arxiv.org/abs/1506.02557v1>

96. Bhutani, K., Sarkar, A., Park, Y., Kellis, M. & Schork, N. J. Modeling prediction error improves power of transcriptome-wide association studies. *bioRxiv* 108316 (2017). doi:10.1101/108316

97. De Jager, P. L. *et al.* A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiology of Aging* **33,** 1017.e1–1017.e15 (2012).

98. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

99. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9,** 179–181 (2012).

100. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS genetics* **5,** e1000529 (2009).

101. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinfor-*

bioRxiv preprint doi: https://doi.org/10.1101/219428; this version posted December 1, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matics **12,** 323 (2011).

102. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).

103. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).

104. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22,** 2008–2017 (2012).

105. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 1–21 (2014).