Title

Causal gene inference by multivariate mediation analysis in Alzheimer's disease

Authors

Yongjin Park[+,1,2], Abhishek K Sarkar[+,1,2,3], 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 (current affiliation)

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

YP: `ypp@csail.mit.edu`; *PLD:* `pld2115@cumc.columbia.edu`; *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 based on imputation of transcriptomic data by utilizing genotypic data and summary statistics to identify putative disease genes cannot distinguish pleiotropy from causal mediation and are limited by overly strong assumptions about the data. To overcome these limitations, we develop a novel Bayesian inference framework, termed Causal Multivariate Mediation within Extended Linkage disequilibrium (CaMMEL), to jointly model multiple mediated and unmediated effects, relying only on summary statistics. We show in simulation unlike existing methods fail to distinguish between mediation and independent direct effects, CaMMEL accurately estimates mediation genes, clearly excluding pleiotropic genes. We applied our method to Alzheimer's disease (AD) and found 21 genes in sub-threshold loci ($p < 1e\text{-}4$) whose expression mediates at least 5% of local genetic variance, and these genes implicate disrupted innate immune pathways in AD.

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 affected by AD[3–5]. The gene regulatory networks of these differentially expressed genes have revealed promising therapeutic targets for ADs[3,4]. However, a fundamental challenge in interpreting case-control differential expression is 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 causal gene within an associated locus. 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]. Regulatory elements are enriched in subthreshold loci ($p < 10\text{-}4$), implicating additional causal variants8–10 and suggesting many causal genes may be distal to the causal variant in GWAS loci.

Here, we present a new method, termed Causal Multivariate Mediation within Extended Linkage disequilibrium (CaMMEL), 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 dimensions of target gene inference problem. (1) Multiple SNPs: CaMMEL leverages multiple SNPs in linkage disequilibrium (LD); (2) multiple genes in tight LD: CaMMEL includes multiple genes in the model and jointly estimate a sparse set of causal genes to avoid spurious associations that may arise due to gene-gene correlations[11]; (3) multiple causal directions: CaMMEL estimates mediation effects adjusting for direct, unmediated effects in expressive modeling framework. We compared statistical power of causal gene identification with two notable methods, summary-based transcriptome-wide associations (sTWAS)[12,13] and Mendelian randomization with Egger regression (MR-Egger)[14], found CaMMEL outperformed in the presence of pleiotropic effects.

We applied CaMMEL to identify mediating genes for AD, combining GWAS of 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)[4,15]. We found 774 protein-coding genes with significant non-zero medication effect (FDR < 1e-4); of them 206 genes are located in proximity of subthreshold GWAS SNPs ($p < 1\text{e-}4$). We focused on 21 mediation genes which explain at least 5% of local genetic variance. These 21 genes include a suicide GWAS gene (*SH3YL1*[16]), peroxisome regulator and cytochrome complex gene (*RHOBTB1*[17–19], *CYP27C1* and *CYP39A1*), and several microglial genes (*LRRC23*[20,21], *ELMO1*[22–24], *RGS17*[25], *CNTFR*[26]) which support functional roles of innate immune responses in disease progression[2,27].

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

Results

Overview of summary-based causal mediation analysis

We propose CaMMEL, a novel method for causal mediation analysis. Our approach jointly models gene expression-mediated (causal) and unmediated (pleiotropic) effects of SNPs on a disease. CaMMEL conceptually extends MR-Egger regression[14] to a multivariate setting[28,29] accounting for multiple correlated SNPs (Fig 1b). Existing methods for TWAS and Mendelian randomiation (MR) ignore direct effects. MR-Egger regression assumes independence between genetic variants and models direct effects as a single intercept term (Fig 1d). To avoid confounded correlation between genes and traits we learned and adjusted for unobserved confounders based on half-sibling regression (Fig 1c; Methods)[30].

Although CaMMEL works on eQTL and GWAS summary statistics, it is easier to think of the method in terms of the underlying generative models (Fig 1d). Assuming that we observed the measurements of a group of mediating variables (i.e., gene expression in this context) for all of the samples, a standard method would fit the data in two

stages. In the first stage, we regress the genotypes on the mediators (eQTL effect size for expression imputation) adjusting for putative reverse causation effects. In the second stage, we regress the effect size for the phenotype against that for the imputed mediators (estimated from eQTL on the mediator) with unmediated genetic effects.

One challenge of this strategy is that 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. To address these limitations, we extended Regression with Summary Statistics (RSS)[31,32] to support multiple summary statistic matrices (effect sizes and standard errors for multiple phenotypes; Methods).

CaMMEL is more powerful than sTWAS and MR-Egger and correctly controls horizontal pleiotropy

We compared CaMMEL with sTWAS[13] and MR-Egger regression[14]. We simulated gene expression data and Gaussian phenotypes using genotypes of 1,709 individuals (Methods). We used 66 real extended loci, randomly sampled 3 per chromosome that 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 (Methods) 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 effects δ (horizontal pleiotropy[33]). 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[34]. 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).

Here, we show results with $n_g = n_q = 2$ as a representative example, but other results with different combinations of n_g, n_q showed similar trends. We found that only with little amount of unmediated effects did sTWAS have comparable power and AUPRC (Fig. 2a) for predicting causal mediating genes. When unmediated effects are spuriously linked to the genes, we found CaMMEL achieved both highest power and accuracy among all competing methods assuming the local PVE by unmediated effects is sufficiently large. By contrast, MR-Egger regression had essentially zero power in all of the simulated scenarios because the SNPs included in the model violates the independence assumption of the method.

To investigate the robustness of methods against infused horizontal pleiotropy, we measured percentage of pleiotropic genes (unmediated) when the methods achieved best precision and recall (maximum F1 score) in causal gene discovery (Fig. 2b). We found almost no evidence for the presence of pleiotropic genes in the discoveries made by CaMMEL (1%) and MR-Egger (2%), but we found pleiotropic genes are more frequently included in the top sTWAS discoveries (5-10%). This partly explains weakened power of TWAS when it confuses mediation with horizontal pleiotropy. In practice, we do not know about how much proportions of gene expression regulatory variances attributes to unwanted pleiotropy and co-regulations[11]. Since TWAS does not promise causal mediation effects[12], rather reports target genes in a broader categories (Fig 1b), we believe our method can provide complementary evidence that improves interpretation of TWAS results.

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 (Methods) containing at least 100 well-imputed SNPs in both data sets. We allowed genes to be assigned to multiple LD blocks (based on SNPs within 1 megabase of the gene body) and empirically calibrated null distribution of mediation by parametric bootstrap[35–38]. We controlled false discovery rate using the method suggested by Storey and Tibshirani[39]. Here, we loosely defined *cis*-regulatory eQTLs and target genes with p-value cutoff ($p < 0.05$) but did not consider genes with no eQTL to

avoid potential weak instrument bias observed in MR on individual-level data[40]. However all GWAS SNPs were included in the background direct effect model if they appeared in the imputed genotype matrix to conservatively estimate mediation effects (by allowing unmediated effects to explain away mediated effects).

We found 774 unique protein-coding genes with significant non-zero show medication effects (FDR < 1e-4), and these genes are located in all the chromosomes without apparent enrichment. Here, we focused on 206 genes whose eQTL SNPs are located in the extended loci with subthreshold GWAS SNPs (p-value < 1e-4; Fig 3a). We found 163 mediating genes outside of GWA significant regions, and of the 43 genes in the genome-wide significant region, we found 9 genes include lead GWAS SNPs among *cis*-eQTLs. In general, a majority of genes explain only a small proportion of genetic variance, less than 1e-4 (Fig 3b), but significantly mediating 774 genes explain on average more than 2% of variance (green line), and the average was approximately same for the 206 genes. However, we found no clear evidence that shows correlation of GWAS significance with the variance explained by mediation effects (Fig 3c) and posterior probability of non-zero mediation effects (Fig 3d). In part, we expect the lack of concordance will be resolved with better powered eQTL summary statistics with appropriate cell type contexts, and joint analysis with different types of molecular QTLs such as DNA methylation will complement the missing fraction of variability.

We prioritized 21 of the 206 genes which explain at least 5% of local genetic variance in the GWAS / subthreshold regions (Fig 3a and Tab 1). We found 8 of the 21 genes are located in genome-wide significant loci (p < 5e-8; Fig 4); two of the high priority genes are included in Fig 4c and Fig 4d. Two cytochrome P450 complex genes, *CYP27C1* (chr2) and *CYP39A1* (chr6) are found significantly expression mediators with permutation p-value essentially zero (p < 3e-6) and explain 15% and 7% of 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 4a). The gene body of *CYP39A1* is also not aligned with cluster of GWAS SNPs, however *cis*-eQTLs are located within LD with the GWAS variants (Fig 4b). Other mediation genes in the GWAS regions such as *DPYSL2* (Fig 4c), *ATP5J2-PTCD1* (Fig 4d) and *ELMO1* (Fig 4e) are also displaced from strong GWAS region, yet *cis*-eQTLs are located within LD with GWAS SNPs. We found known AD genes, *COPS6* (Fig 4c), *CLU* (Fig 4d) and *MS4A4A* (Fig 4f), are significant mediators and co-localized with GWAS signals.

Many of these genes highlight dysfunctional innate immunities and resulting stress responses in AD. Molecular function and disease-causing mechanism of *CYP27C1* yet to be elucidated[41], but *CYP39A1* is involved in the generation of 25-hydroxy cholesterol from cholesterol, which then regulates innate immunity[42]. A murine experiment showed that inflammation in central nervous system due to disrupted lipid metabolism down-regulated activities of cytochrome P450 complex[43]. *DPYSL2* was up-regulated in cortex, striatum and hippocampus after ischemic stroke[44] and was listed among genes modulate activiation of microglial cells[45]. *ELMO1* confers risk of immune disorders such as multiple sclerosis in previous GWAS[23,24], and triggers phagocytosis in microglia interacting with other proteins[22].

However, we note several case where interpretation of CaMMEL results can be little unclear. For instance, *cis*-eQTLs of *CLP1* are located almost independent of the significant GWAS region on chromosome 11, but this gene explains 11% of local variance (Fig 4f). Even after we extend effective boundaries the *cis*-eQTLs through LD, effects of *CLP1* seem outside of the highest GWAS peak in the region. Nonetheless, *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[46,47].

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 5a). *SH3YL1* was found significant in GWAS with attempted suicide and expressed in brain, but underlying mechanisms are unknown[16]. Interestingly this gene was also found robustly associated height in previous sTWAS[12,13]. Several previous researches noted subtle relationship between AD and suicide attempt[48,49] and height[50]. 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 5b). 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 response to optic neuronal injury[20] and enriched in ramified microglial cells[51]. *CLSTN3* (calsyntenin-3) interacts with neurexin and forms synpatic adhesion[52] and up-regulated by increase of amyloid beta protein and accelerated neuronal death[53].

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 5c). 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[26]. 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[54], and expression of *EXOSC3* was differentially regulated in human monocyte and macrophage induced by lipopolysaccharide, suggesting functional roles in inflammatory bowel disease[55].

Discussion

We developed novel model-based approach to causal mediation analysis that infers regulatory mechanisms from summary statistics of GWAS and eQTLs. Our approach borrows great ideas from two existing methods, MR-Egger regression and TWAS: Key idea of MR-Egger is to distinguish mediated effects from unmediated ones[14]. We formulae them in a Bayesian framework based on well established summary statistics models, but eliminates necessity of explicit fine-mapping of SNPs to account for LD structure[56]. Instead, we automatically carried out variable (mediator) selection through spike-slab prior[57] on the mediation effect sizes, by which we avoid unwanted collinear between mediators[58,59]. The methodology is generally applicable to general regression or summary-based regression problems in genetics and regulatory genomics.

Recent methods for TWAS have made important advances in identifying genes which are more likely to be causal. Key idea of TWAS is to properly integrate evidences of multiple SNPs while taking into account of local LD structure[12]. However, TWAS cannot distinguish between causal mediation, pleiotropy, linkage between causal variants, and reverse causation, which could lead to inflated false positives as pleiotropy is quite prevalent in genetics[60].

Univariate MR[61] has been applied to complex traits using genotype data[62,63] and summary statistics[56] from separate studies. Unlike naive MR requires strong assumption of no pleiotropic effects, horizontal pleiotropy[33,61] (the SNP affects both genes and phenotypes) or LD-level linkage (two causal SNPs affecting genes and phenotypes in LD)[60], MR-Egger meta-analyzes the estimated causal effect sizes across multiple SNPs using Egger regression and provides unbiased estimation of directional pleiotropy[14], with a strong assumption that SNPs used as IVs should be uncorrelated and directional pleiotropy can be estimated by a single scalar value.

Previous MR analysis on DNA methylation[64] used stringent FWER cutoff on instrument variable selection to avoid weak instrument bias observed in MR on individual-level data . Here we retained *cis*-regulatory eQTLs and target genes with lenient p-value cutoff (0.05) for several reasons: (1) CaMMEL estimates RSS model with in-sample LD covariance, while reducing the risk of double counting evidences from multiple weak IVs, and a small number of variants would be truly effective with spike and slab prior. (2) We expect that as sample sizes increase, every gene will be an eGene, and then we would include every gene even with a stringent cutoff; (3) SNPs can be strong QTLs but have no indirect effect on phenotype; (4) analyzing multiple mediators allows the causal one to explain away the others. Idea of joint analysis on multiple genes is to make genes compete with each other in local regions. Elimination of the potential competitors would over-emphasize genes with strong eQTLs effects although causal eQTL variants are not necessarily lead SNPs[23,65,66].

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[67].

Methods

Causal mediation model specification

For each SNP j we would have summary effect size $\hat{\theta}_j$ and variance $\hat{\sigma}_j^2$. The RSS model provides a principled way to describe a generative model of the GWAS summary data. For p number of SNPs, a summary effect size vector $\hat{\boldsymbol{\theta}} \equiv (\hat{\theta}_1, \ldots, \hat{\theta}_p)^\top$ can be generated from multivariate Gaussian model while taking into account dependency structure of $p \times p$ linkage disequilibrium (LD) matrix R:

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

Here our interest is on "true" (multivariate) effect size vector $\boldsymbol{\theta}$ and we abbreviated a diagonal matrix of expected squares, S, i.e., $S_{jj} = \sqrt{\mathbb{E}[\hat{\theta}_j^2]} = \sqrt{\hat{\theta}_j^2/n + \sigma_j^2}$ with appropriate GWAS sample size n.

In practice we estimate $\hat{R}$ from standardized $n \times p$ genotype matrix G of n individuals by crossproduct, for instance $\hat{R} = G^\top G/(n-1)$. We could regularize R matrix by adding ridge regression diagonal matrix (e.g.,[68]), or by dropping variance components corresponding to small eigen values (e.g.,[69]).

Mediation analysis aims to decompose total effect size vector $\boldsymbol{\theta}$ into direct and mediated effects. For each mediator $k \in [K]$, there are multivariate SNP effects $\boldsymbol{\alpha}_k$, additionally genetic effects on this mediator is propagated to GWAS trait with medication coefficient β_k, but there is also un-mediated direct effect $\boldsymbol{\gamma}$: $\boldsymbol{\theta} = \sum_{k=1}^K \boldsymbol{\alpha}_k \beta_k + \boldsymbol{\gamma}$.

In previous analysis based on MR and sTWAS methods, hypothesis test focused on single mediator variable at a time treating them independently, i.e., $K = 1$. Assuming GWAS marginal effects $\hat{\boldsymbol{\theta}}$ and QTL effects on mediators $\hat{\boldsymbol{\alpha}}$ follow the RSS model, therefore $\mathbb{E}[\hat{\boldsymbol{\theta}}] = SRS^{-1}(\sum_k \boldsymbol{\alpha}_k \beta_k + \boldsymbol{\gamma})$ and $\mathbb{E}[\hat{\boldsymbol{\alpha}}_k] = S_\alpha R S_\alpha^{-1} \boldsymbol{\alpha}_k$. We can reasonably model $\mathbb{E}[\hat{\boldsymbol{\theta}}] \approx \sum_k \hat{\boldsymbol{\alpha}}_k \beta_k + SRS^{-1}\boldsymbol{\gamma}$ with no measurement error assumption.

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

Here we assume standard errors of GWAS effects can be substituted with standard errors of QTL effects up to some scaling factor, because association statistic is mainly determined by underlying minor allele frequencies and samples sizes, and the scaling factor was clumped with mediation parameter.

Model inference

We established independent LD blocks of SNPs within a chromosome and restricted 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. We then merged LD blocks overlapping in genome.

The key challenge in fitting the causal mediation model is dealing with the covariance matrix in the likelihood. To address this challenge, we exploit the spectral decomposition of the LD matrix (singular value decomposition of the genotypes), that is $(n)^{-1/2}G = UDV^\top$, therefore LD matrix $R = VD^2V^\top$.

Defining $\tilde{G} \equiv V^\top$ and $\tilde{\mathbf{y}}_t \equiv V^\top S^{-1}\hat{\boldsymbol{\theta}}_t$ we obtain equivalent, but fully factorized, multivariate Gaussian model:

$$\mathbf{y}_t \sim \mathcal{N}\Big(D^2\tilde{G}S^{-1}\boldsymbol{\theta}_t, D^2\Big)$$

We assume the spike-and-slab prior on each vector of effect sizes $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ in order to select between correlated SNPs/mediators[57,70]. We efficiently fit the models using black box variational inference[71,72]. We additionally exploit the spectral transformation of the LD matrix to make the model inference more amenable.

Further, letting $\eta_k \equiv \sum_{j=1}^{p} V_{jk}S_j^{-1}\theta_j$, we can rewrite the transformed log-likelihood of the model for each eigen component i:

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

This reformulation permits efficient and stable stochastic gradient updates in combination with control variate techniques[67].

Half-sibling regression

We learn and correct for unobserved confounders based on half-sibling regression (Fig 1c)[30]. For each gene we find 105 potential control genes, selecting most correlated 5 genes from each other chromosome. We assume *cis*-regulatory effects only occur within each chromosome and other inter-chromosomal correlations are due to shared non-genetic confounders. Further, to ensure not to eliminate *cis*-mediated *trans*-effect, 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.

Parametric bootstrap

Variational inference only captures 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[35,36,38].

Simulation framework

We used a generative scheme proposed by Mancuso *et al.*[13] with a slight modification that includes direct (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 direct (un-mediated) 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 direct 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 $\epsilon \sim \mathcal{N}(0, 1-\rho)$.

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, $\beta \sim \mathcal{N}(0, \tau/n_g)$, $\gamma \sim \mathcal{N}(0, \delta/n_d)$ and $\epsilon \sim \mathcal{N}(0, 1-\tau-\delta)$.

We compared two existing methods: MR-Egger regression[14] and summary-based TWAS[12,13].

Calculation of TWAS test statistics from summary z-scores

For simplicity we used completely summary-based approach[13] 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, $\boldsymbol{\alpha} \approx R^{-1}\mathbf{z}_T/\sqrt{n}$.

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

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

Calculation of proportion of variance explained

We measured explained variance according to the definition of Shi *et al.*[69] Total variance can be decomposed into (1) mediated component, $(\boldsymbol{\alpha\beta})^\top R(\boldsymbol{\alpha\beta})$ and (2) un-mediated (direct) 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.

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)[73]. 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[74] using `SHAPEIT`[75], retaining only SNPs with MAF > 0.05. We then imputed SNPs in 5MB windows using `IMPUTE2`[76] 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`[77] and we focused on 18,461 coding genes out of 55,889 according to the GENCODE annotations (v19 used in the GTEx v6p, in press).

Original gene-level quantification data follows negative binomial distribution. We adjusted variability of sequencing depth across samples following previous methods[78,79]. 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[80]. 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.

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.

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 < 1e-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

Figures

Figure 1. Causal mediation analysis redeems genuine mediation effect. **(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 be interpreted in multiple modes: mediation, pleiotropy and reverse causation. **(c)** Half-sibling regression[30] identifies hidden confounder of gene expression using control genes found in different chromosomes. **(d)** Causal mediation analysis jointly estimate two regression models taking into accounts of multiple sources of gene expression variation.

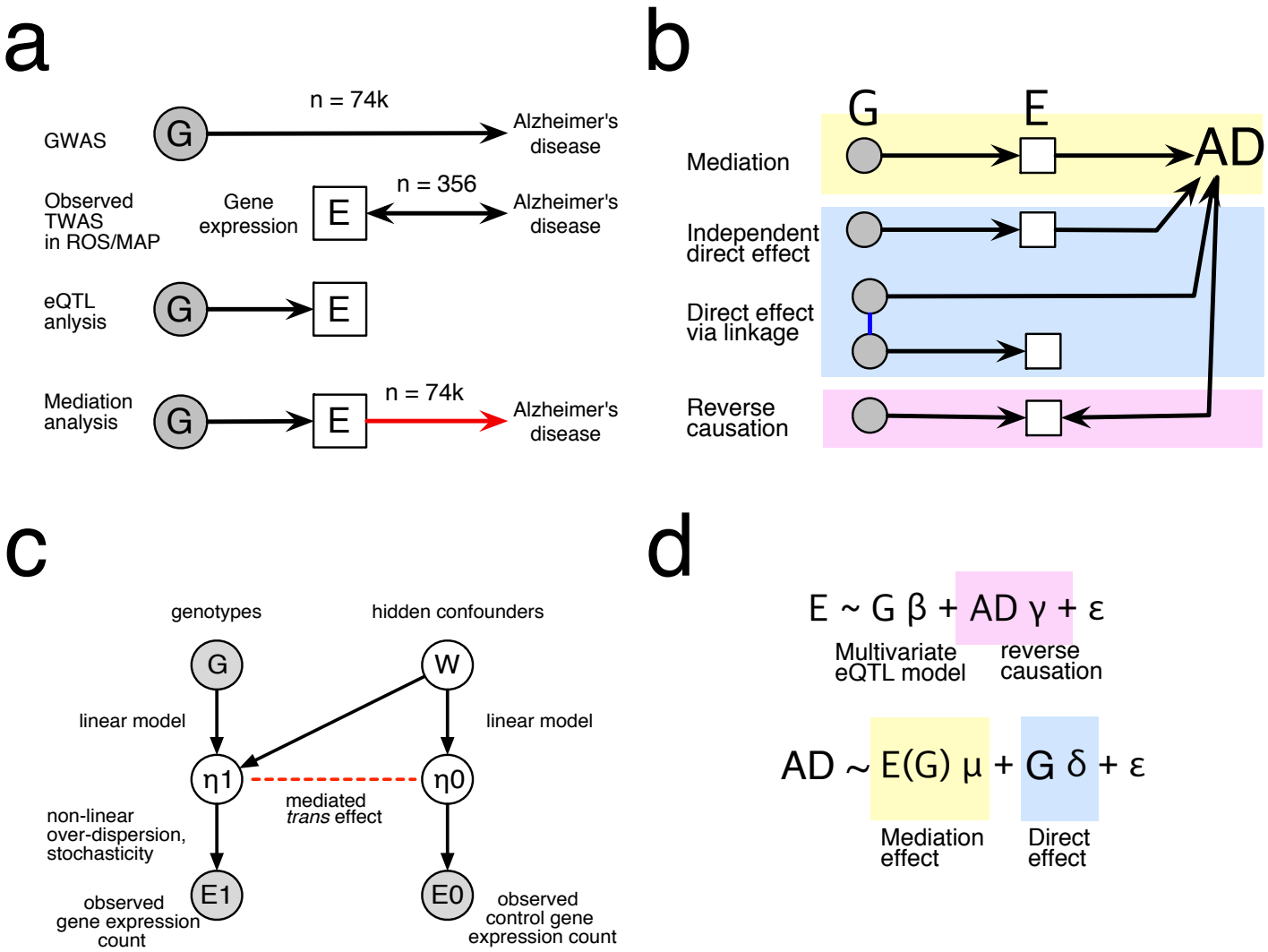

Figure 1: Overview of study design and methods

Figure 2. Simulation studies show superior statistical power of CaMMEL over other methods without inflated false positives due to horizontal pleiotropy. *CaMMEL*: causal multivariate mediation within extended linkage disequilibrium (this work); *TWAS*: summary-based transcriptome wide association[13]; *MR-Egger*: Meta-analysis of GWAS and eQTL summary statistics by Egger regression[14]; *AUPRC*: area under precision recall curve; *F1*: harmonic mean of recall and precision; *FDR*: false discovery rate. **(a)** Statistical power and AUPRC varying degree of spurious direct effects. **(b)** Percentage of pleiotropic genes at best precision-recall.

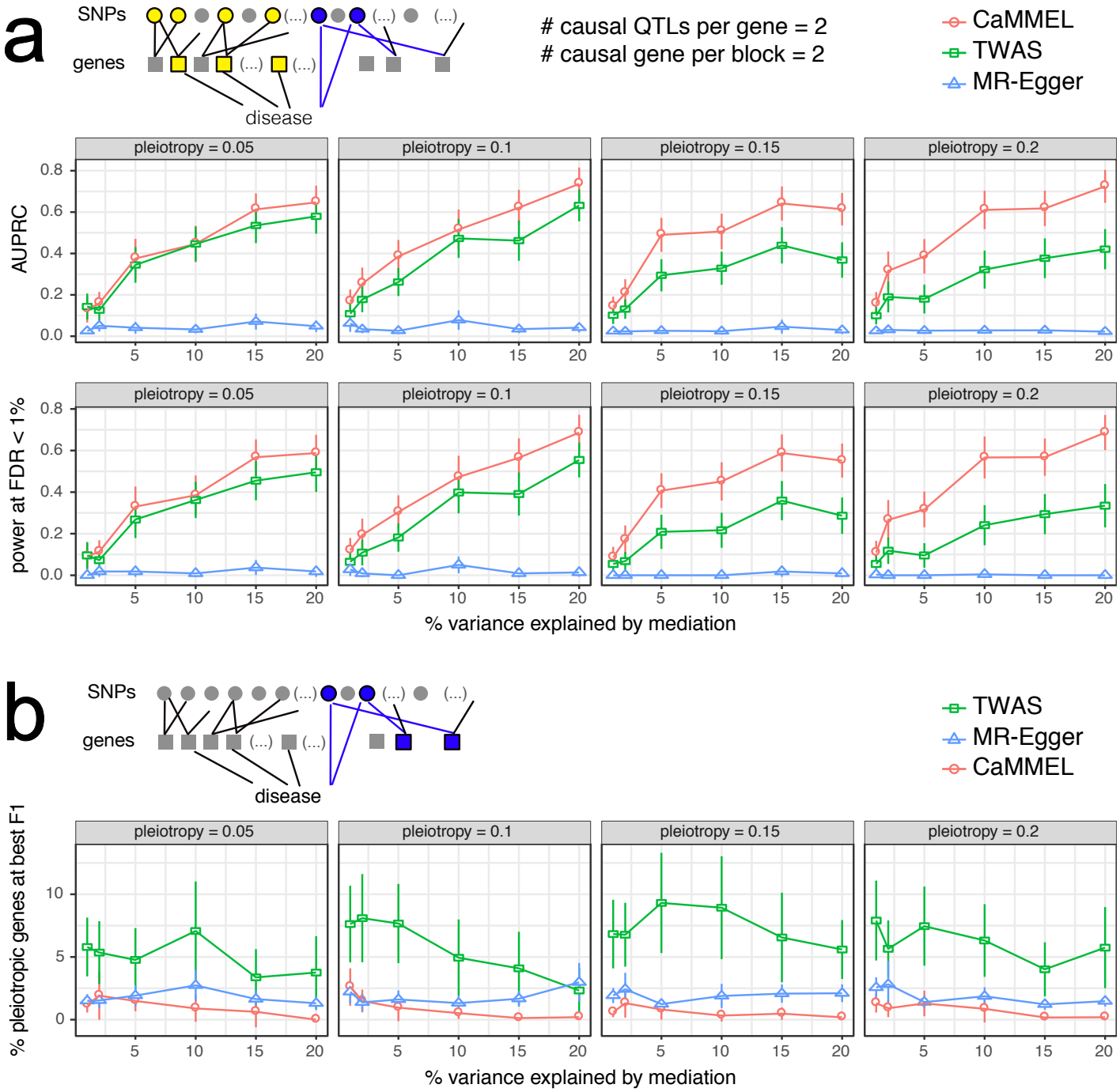

Figure 2: Simulation results using actual ROSMAP genotype matrix

Figure 3. Transcriptome-wide mediation analysis reveals causal target genes in AD. **(a)** Mediation effect sizes over all the genes located in GWAS subthreshold region (p < 1e-4). Size of dots are proportional to the proportion of local genetic variance explained by mediation. Errorbars show 95% confidence intervals. Dots are colored differently according to the GWAS significance. 21 genes (PVE > 5%) are annotated (see the text). **(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 < 1e-4); *blue line*: histogram of PVE for 206 significant mediation genes found in GWAS subthreshold regions (FDR < 1e-4, GWAS p < 1e-4). *Green horizontal line*: average PVE for 774 genes. *Red horizontal line*: 5% cutoff. **(c)** Density estimation showing relationship between PVE (y-axis) and GWAS significance level (x-axis). We marked 21 strong genes with 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 < 1e-4 cutoff. **(e)** Density estimation showing relationship between best eQTL (y-axis) and best GWAS significance level (x-axis) within LD.

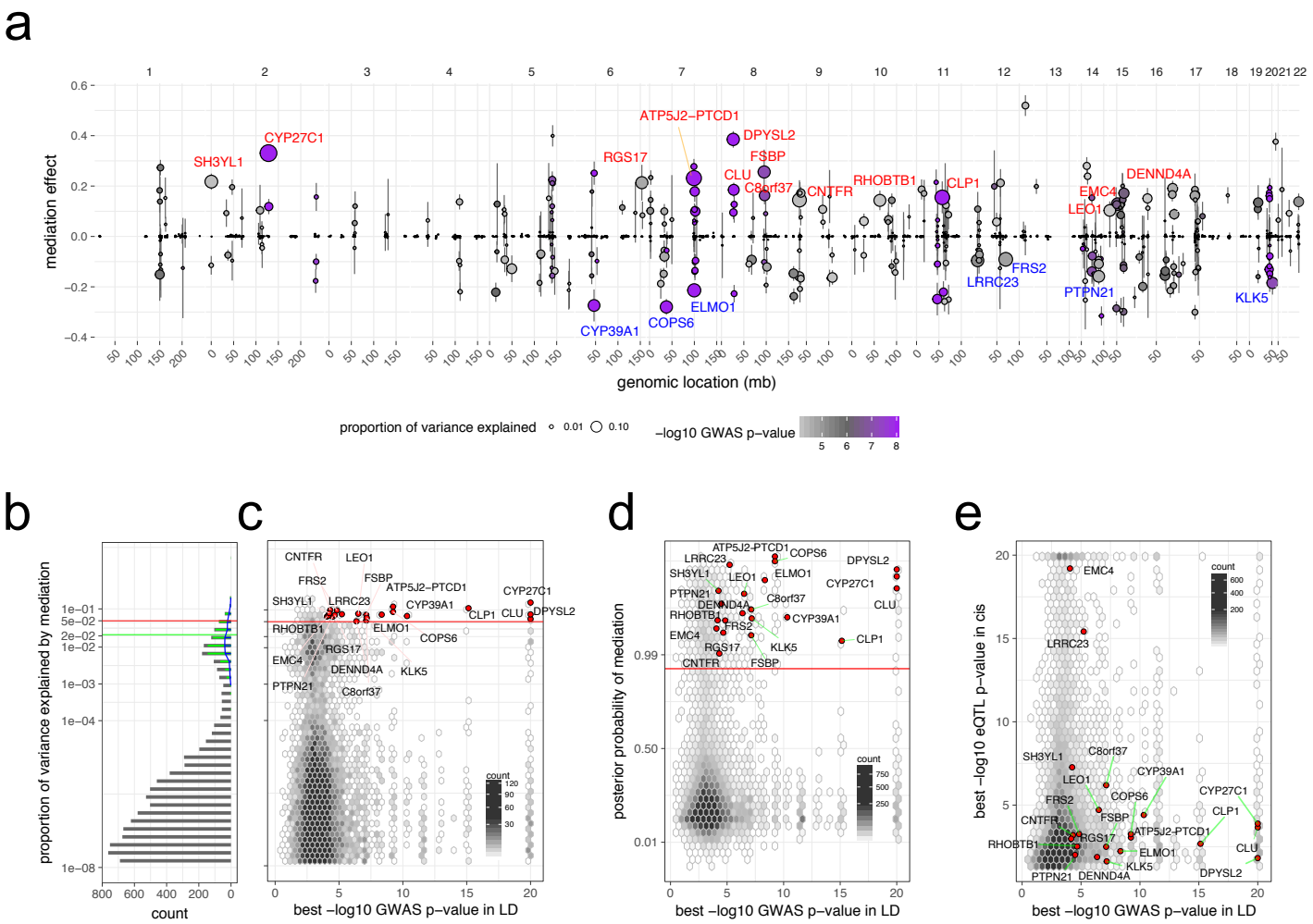

Figure 3: Genome-wide patterns of gene expression mediation in AD.

Figure 4. 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.

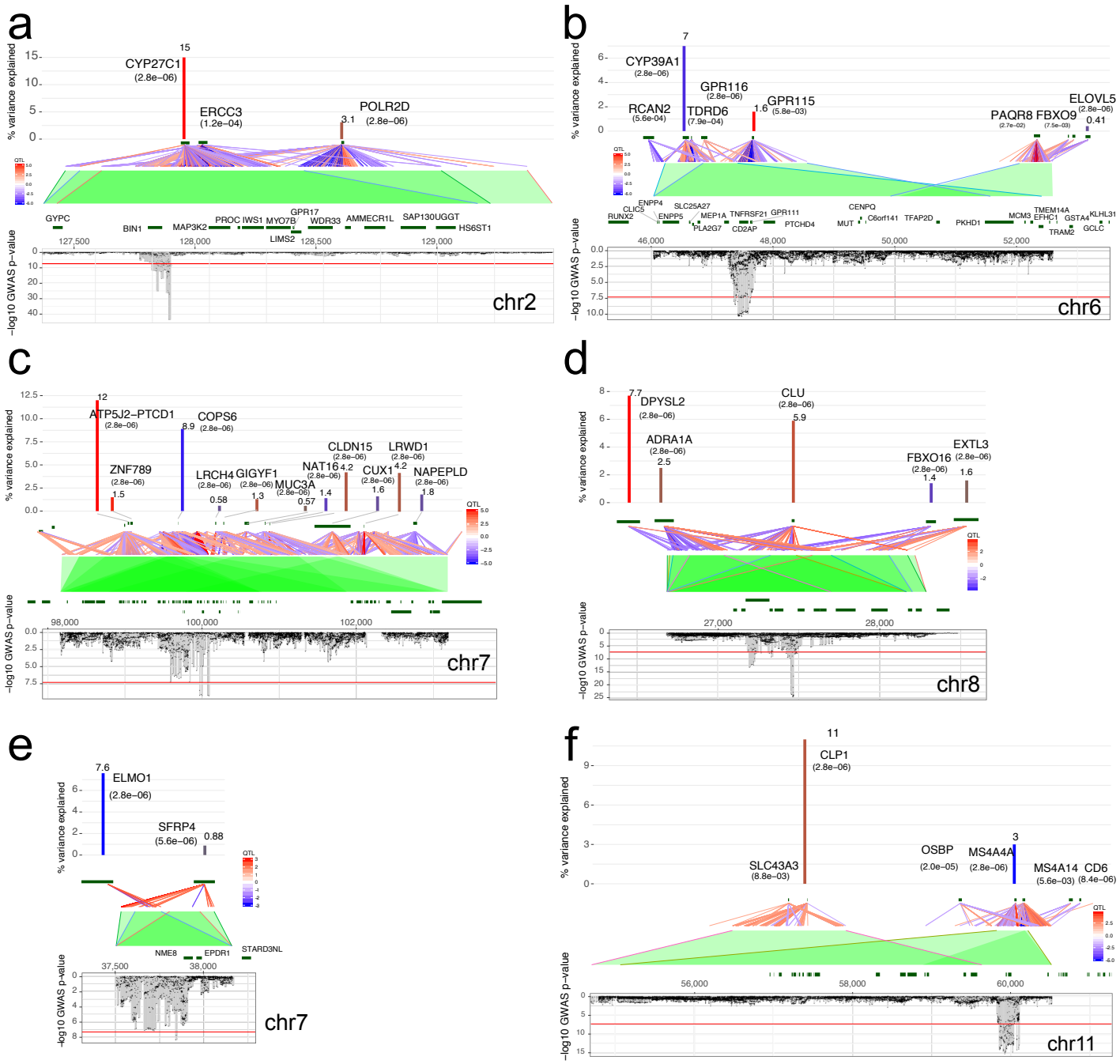

Figure 4: Examples in the GWAS regions

Figure 5. 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.

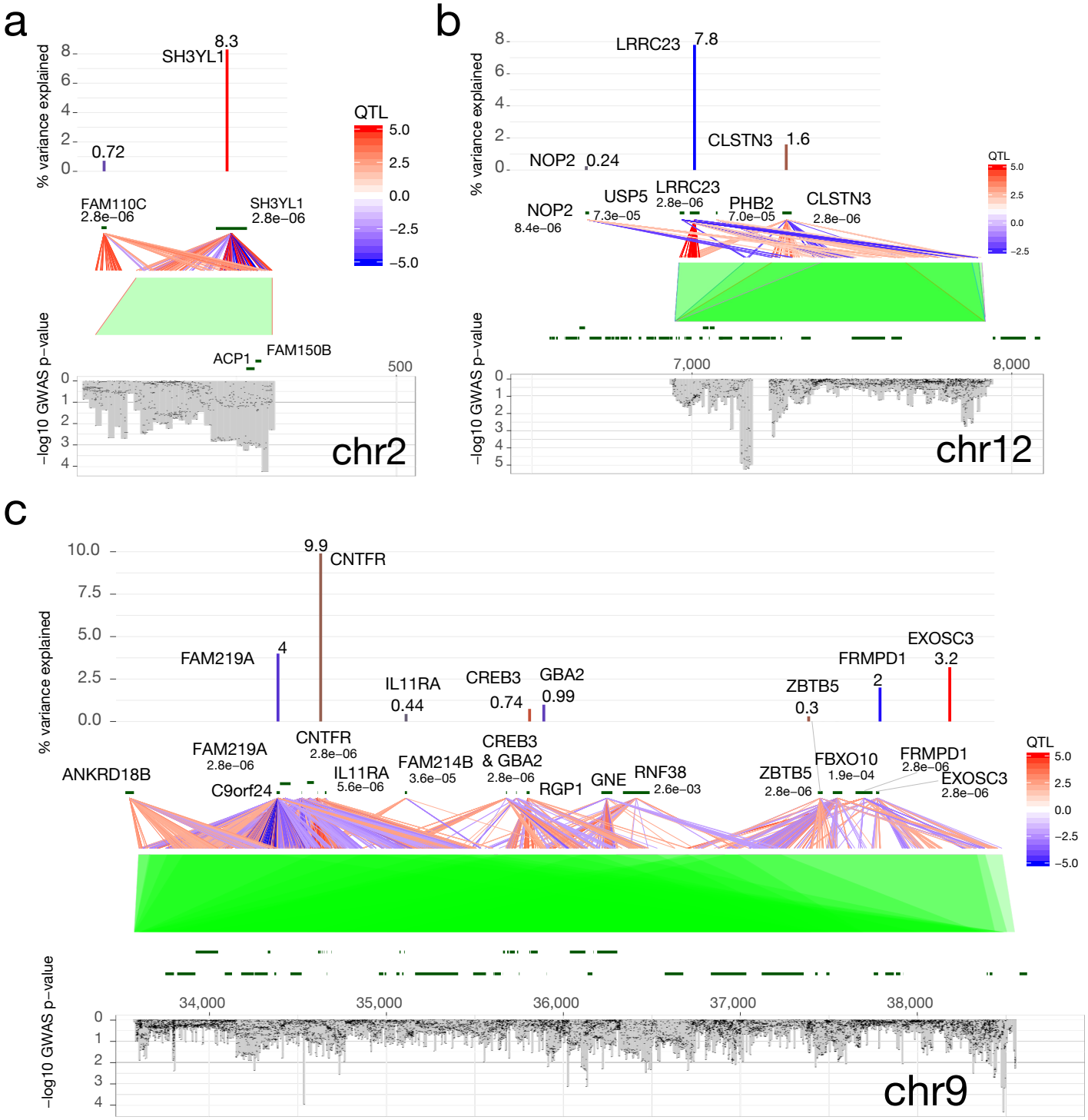

Figure 5: Examples in the sub-threshold regions

Acknowledgment

We thank Alkes Price (Harvard) and Bogdan Pasaniuc (UCLA) for insightful discussion.

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. AKS and LH contributed key ideas in method development. YP implemented `C++` and `R` software. YP carried out mediation analysis. JD helped network analysis. YP, AKS, LH, 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 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. Wainberg, M. *et al.* Vulnerabilities of transcriptome-wide association studies. *bioRxiv* 206961 (2017). doi:10.1101/206961

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

13. 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).

14. 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).

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

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

17. 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).

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

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

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

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

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

23. 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).

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

25. 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).

26. 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).

27. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimers disease. *Nature* **518,** 365–369 (2015).

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

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

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

31. 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).

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

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

34. 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

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

36. 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).

37. 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).

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

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

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

41. Guengerich, F. P., Cheng, Q. & Ma, Q. Orphans in the Human Cytochrome P450 Superfamily: Approaches to Discovering Functions and Relevance in Pharmacology. *Pharmacological Reviews* **63,** 684–699 (2011).

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

43. 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).

44. 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).

45. 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).

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

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

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

49. 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).

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

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

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

53. 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).

54. 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).

55. Baillie, J. K. *et al.* Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease. *PLoS genetics* **13,** e1006641 (2017).

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

57. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical . . .* **83,** 1023–1032

(1988).

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

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

60. 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).

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

62. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27,** 1133–1163 (2008).

63. Hinke Kessler Scholder, S. von, Smith, G. D., Lawlor, D. A., Propper, C. & Windmeijer, F. Mendelian randomization: the use of genes in instrumental variable analyses. *Health Economics* **20,** 893–896 (2011).

64. 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

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

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

67. 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

68. 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).

69. 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).

70. 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).

71. 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>

72. 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>

73. 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).

74. Consortium, T. 1. G. P. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

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

76. 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).

77. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12,** 323 (2011).

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

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

80. 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).