Title

Regulatory landscape of Alzheimer's disease variants in human microglia

Authors

Chia-Yi Lee[1,2*], Ashvin Ravi[3-6*], Tzu-Chieh Huang[1,2], Briana Wyman[1,2], Ariana Chriss[3-6], Jack Humphrey[3-6], David A Knowles[7,8], Chris Cotsapas[1,9], Laura Huckins[2#], Towfique Raj[3-6#], Kristen Brennand[1,2,4,5#]

Affiliations

[1] Department of Genetics, Wu Tsai Institute, Yale University School of Medicine, New Haven, CT, USA
[2] Department of Psychiatry, Division of Molecular Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[3] Ronald M. Loeb Center for Alzheimer's Disease, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA
[5] Nash Family Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6] Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[7] New York Genome Center, New York, NY, USA
[8] Department of Computer Science, Columbia University, New York, NY, USA
[9] Department of Neurology, Yale University School of Medicine, New Haven, CT, USA

[*]Contributed equally
[#]Co-corresponence: kristen.brennand@yale.edu, towfique.raj@mssm.edu and laura.huckins@yale.edu

Abstract

Although defined by end-stage neuronal loss, neurodegenerative diseases are increasingly hypothesized to be initiated by early microglial dysfunction. Genetic risk for Alzheimer's disease (AD) is particularly highly enriched for non-coding variants located in microglia-specific regulatory regions. Yet, the linkage-dense nature of these variants limits their functional interpretation. To elucidate how risk variants influence microglial gene regulation, we contrast fine-mapping of AD genome-wide association study (GWAS) loci by four complementary methods, and resolve a highly concordant set of putative causal variants. In total, 11,550 candidate regulatory sequences (CRSs), nominated from GWAS (n = 2,183), microglial expression quantitative trait loci (QTL) (n = 5,721) and microglial chromatin accessibility QTL (n = 1,519), are screened *en masse* by massively parallel reporter assay (MPRA) in human induced pluripotent stem cell (hiPSC)-derived microglia (iMGL) and immortalized HMC3 cell lines. Distinct sets of functional variants are identified between cellular models, underscoring the importance of using physiologically relevant microglia for variant interpretation. Functional regulatory variants (frVar) are enriched in promoters and splicing regions, with over half predicted to disrupt transcription factor (TF) binding. The affected TFs yield distinct regulatory modules, the first with an inflammatory program and the second with a DNA stress-response network. Mapping frVar to their predicted target genes reveals enrichment for biological processes related to amyloid biology, immune signaling, and neuron-microglia communication. Notably, we pinpoint functional variants within several understudied AD loci, including *TSPOAP1*, providing direct evidence for causal regulatory effects. Together, these findings establish the first large-scale microglial-context MPRA, define a microglial regulome underlying AD genetic risk, and provide a functional framework to refine noncoding variant interpretation and therapeutic target discovery.

Introduction

Alzheimer's Disease (AD) is the leading cause of dementia in the elderly. It is driven by a complex interaction between age-related neurobiological processes and genetic risk (twin study heritability estimates range from 58%-79%[1]). Recent genome-wide association studies (GWAS) have associated 75 independent loci with genetic risk for AD[2]. However, the majority of these variants map to non-coding regions of the genome[3], so that pinpointing functional variants and their molecular effects remains challenging. In particular, non-functional variants in linkage disequilibrium (LD) with true functional alleles are only partially resolved by statistical approaches that incorporate genomic annotations and LD reference panels[4]. In fact, the most recent fine-mapping effort nominated only 21 putative causal variants[5].

GWAS variants frequently localize to cell-type-specific regulatory regions[6]; in AD, they are highly enriched for microglia enhancer regions[7,8]. Consistent with this, microglia play a key role in AD progression[9,10]. AD GWAS signals can be mapped to putative functional roles by colocalization with microglia gene expression quantitative trait loci (eQTL)[11,12] or chromatin accessibility quantitative trait loci (caQTL)[7], uncovering novel gene-disease and peak-disease associations. Yet statistical inference alone has proven insufficient to comprehensively resolve all of the functional variants and regulatory mechanisms through which genetic risk may contribute to AD[7,13].

Massively parallel reporter assays (MPRA) can bridge this gap by systematically and simultaneously testing allelic effects, pairing each variant with a reporter gene bearing unique transcribed barcodes, thereby measuring the regulatory activity of thousands of variants in parallel by high-throughput sequencing[14,15]. To date, most MPRA have been conducted in transformed cell lines[16–18], with uncertain relevance to human brain cell contexts[18]. Moreover, our recent attempt to map neurodegenerative disease and psychiatric disorder GWAS variants resolved minimal functional activity of AD GWAS loci in human induced pluripotent stem cell (hiPSC)-derived neurons[19], reinforcing the need to test regulatory mechanisms driving AD risk specifically in human microglia.

Here, we performed fine-mapping of the most recent AD GWAS to prioritize functional variants and integrated these with microglial eQTL and caQTL maps, in total nominating 11,550 risk variants. By conducting MPRA in parallel in immortalized human microglia (HMC3) and human induced pluripotent stem cell (hiPSC)-derived microglia-like cells (iMGL), we conducted the first systematic effort to functionally interrogate Alzheimer's- and microglia-associated non-coding variants. We tested whether specific variant prioritization approaches (e.g., fine-mapping, eQTL, caQTL) more effectively predicted MPRA activity and if distinct cellular models (HMC3 and iMGL) more effectively resolved regulatory activity. Our variant-level, gene-level, and network-level analyses decipher the regulatory outcome of these functional variants and link genetic risk to cell-type-specific regulatory mechanisms in AD. Altogether, we demonstrate that integrating statistical genetics with high-throughput experimental validation clarifies GWAS ambiguity and provides a framework for functional studies in aging-related neurodegeneration.

Results

Fine-mapping prioritized functional variants enriched in microglia enhancers

Statistical fine-mapping refined AD GWAS association signals and prioritized functional variants (computational schematic, **Fig. 1**). Four widely used Bayesian and functionally informed methods (SuSiE[20,21], FINEMAP[22], PolyFun[23]+SuSiE, and PolyFun+FINEMAP) were evaluated to assess the robustness of fine-mapping results across approaches. Results were highly concordant: 53 of 75 AD risk loci harbored at least one fine-mapped variant (posterior inclusion

probability (PIP) > 0.1) across all four methods (95% credible sets). The distribution of credible sets revealed important distinctions in the strength of fine-mapping resolution: loci with larger credible sets were frequently composed of variants with low PIP that were less likely to contain functional variants, whereas loci with smaller credible sets were enriched for high-PIP variants with stronger support for specific functional candidates (**Fig. 2A)**. A substantial fraction of variants were consistently prioritized across all fine-mapping methods (**Fig. 2B**). Our analysis recaptured the well-characterized *TREM2* R47H missense variant (rs75932628)[24,25] as well as novel noncoding variants at the locus, whereas at the *BIN1* locus, rs6733839[6,26,27] emerged as the sole credible functional variant (**Fig. 2C**). Intersection of the 95% credible set variants with cell type–resolved regulatory annotations revealed that fine-mapped variants were disproportionately located in microglial enhancers (**Fig. 2D; Supplementary Fig. 1A**) and particularly pronounced at AD-associated loci implicated in immune regulation and microglial biology (e.g., *PLCG2, SPI1,* and *RASGEF1C*). As the PIP threshold increased, the relative fraction of variants mapping to microglial enhancers remained consistently greater than for other brain cell types (**Fig. 2E**). Similarly, fine-mapped AD variants were significantly enriched in microglial enhancers relative to other immune cell types[28] (**Fig. 2F**), a pattern not observed for bipolar disorder **(Supplementary Fig. 1B)**. Fine-mapped variants demonstrated stronger enrichment for microglial enhancers than lead AD GWAS variants (**Fig. 2G; Supplementary Fig. 1C**). Altogether, fine-mapping refined AD GWAS signals and enhanced the resolution of microglia-specific functional interpretation.

hiPSC-derived microglia, more so than HMC3s, capture AD-relevant regulatory programs

A comprehensive set of candidate AD-associated functional variants were prioritized for MPRA assessment. First, all AD fine-mapped variants were included as well as those identified in the most recent large-scale AD fine-mapping study (PIIP > 0.01; total variants = 2,183). Second, lead microglia eQTL variants (n = 5,721) from a meta-analysis of 391 donors[12] were incorporated, along with variants in strong LD ($r^2 > 0.8$) with these eQTLs. These were further filtered to retain only those overlapping either significant microglia ATAC-seq peaks (n = 1,519) or microglia ABC enhancer–gene links (n = 730). Third, fine-mapped microglia caQTL variants with high confidence fine-mapping (PIP > 0.5; n = 2,074) were included[7]. To anchor our set with robust association signals, we added lead AD GWAS variants from the three most recent genome-wide studies (n = 135)[2,5,29,30]. Together, this integrative approach yielded a consensus set of 11,550 putative functional variants, representing the most credible regulatory candidates for AD. For testing both alleles of each variant, we synthesized and tested 23,100 candidate regulatory sequences (CRSs), along with 844 scrambled and positive/negative controls (experimental schematic, **Fig. 1**).

In our MPRA library, each CRS was linked to a unique barcode (mean = 119, median = 106 barcodes per CRS), and the abundance of barcode RNA relative to its corresponding DNA [$\log_2$(RNA/DNA)] was used to quantify transcriptional activity. The MPRA library was transduced into mature iMGLs (40 DIV; two control donors, three replicates from a male and four from a female donor) or into HMC3 cells 72 hours before harvest (three biological replicates each). The assay achieved high coverage, with approximately 95% of CRSs detected in both microglial models and strong inter-replicate correlations among DNA counts, RNA counts, RNA/DNA ratios (r = 0.60-0.88), and barcode counts (**Supplementary Fig. 2**). To ensure reliable activity estimates, CRSs with <10 barcodes or with low RNA/DNA ratio were excluded (**Fig. 3A**). Functional regulatory variants (frVar) were defined as sites with significant allele-specific activity ($p_{FDR} < 0.05$).

Surprisingly, the method of variant prioritization, whether GWAS or QTL-based, did not predict CRS activity or the proportion of frVars in MPRA. In both HMC3 and iMGLs, the fraction

of active CRSs was not enriched between fine-mapped GWAS credible sets, lead eQTLs, caQTLs, or their LD proxies (**Fig. 3B**). Likewise, whether nominated by GWAS or eQTL methods, the likelihood that an frVar exhibited transcriptional activity or allelic regulatory effects was equal (**Supplementary Fig. 3A**). Although multiple nominations (i.e., two or more of GWAS, eQTL, and caQTL) was a significant positive predictor of frVar status ($\beta = 0.18$, $p = 0.03$; **Supplementary Fig. 3B-C**), only ~6% of the library carried more than one annotation, suggesting that these datasets capture largely distinct subsets of functional variants with modest convergence.

In total, 2,003 frVar were identified in iMGL and 1,181 in HMC3 (**Fig. 3C**). Unexpectedly, there was very little overlap in frVar between cell types: only 21.5% of HMC3 signal replicated in iMGLs (**Fig. 3D**). As iMGLs exhibited the strongest regulatory signals and a greater number of frVar than HMC3, we focused downstream analyses on iMGL. This choice was reinforced by the greater transcriptomic similarities of iMGLs to primary microglia.

Transcriptomic analyses revealed substantial differences between iMGL and HMC3[31]. Principal component analysis (PCA) highlighted broad differences (**Supplementary Fig. 4A**). Relative to primary microglia and iMGL, HMC3 cells more robustly expressed cell cycle programs (chromosome segregation, nuclear division, and mitosis) and less well expressed immune and metabolic pathways (immune cell activation and immune response) (**Supplementary Fig. 4B,C**). Gene Ontology (GO) enrichment of nearest genes to frVar revealed that iMGL were more enriched for hallmarks of microglial function and neuron-microglia communication, such as synapse organization, regulation of cell shape and adhesion, and calcium-mediated signaling pathways (**Fig. 3E, left**). In contrast, HMC3 enrichments emphasized generic signaling and metabolic programs (**Fig. 3E, right**). Although the overlapping signals indeed centered around immunological synapse formation (**Fig. 3E, middle**), the largely divergent regulatory programs between iMGLs and HMC3 underscore the importance of assaying variant function in a physiologically relevant microglial model.

iMGL frVar concentrate in promoter-proximal features and regulate disease-relevant TF

Fine-mapping improved the likelihood of detecting MPRA positive variants, with high posterior inclusion probability (PIP) variants ($PIP \geq 0.5$) more likely to be iMGL frVar compared to low-PIP variants (low $0.1 \leq PIP < 0.2$) (pairwise Wilcoxon test, BH-adjusted $p = 0.018$) (**Fig. 4A**). iMGL frVar were significantly enriched in canonical regulatory elements, particularly ENCODE promoter and 5'UTR regions (**Fig. 4B**). To further evaluate cell-type-specific activity, we compared MPRA-positive (frVar, $p_{FDR} < 0.05$) versus MPRA-negative ($p_{FDR} > 0.05$) sequence and found significant enrichment of MPRA hits in microglial promoter regions ($OR = 1.28$, $p = 0.02$) (**Fig. 4C**). Similar enrichment was observed in iMGL ATAC-seq peaks under a more stringent MPRA FDR threshold (**Supplementary Fig. 5**).

Consistent with promoter-proximal regulation, ~50% of CRSs overlapped known transcription factor (TF) binding sites (**Fig. 4E**). For example, the high confidence frVar rs147895729 showed distinct transcriptional activity difference between the reference and alternative alleles, which were predicted to disrupt *FOXO1* and NKX2.2 binding (**Fig. 4D**). Top enriched TFs reveal links to neurodegeneration[32–34] and neurodevelopment[35–39], notably SPI1, a key myeloid regulator known to be enriched among AD risk variants[7,11] and to modulate amyloid-β pathology[40,41]. To better understand how these TFs may regulate microglial gene programs, we constructed a microglial gene regulatory network (random-forest approach[42]) using co-expression between TF and target gene expression profiles (**Fig. 4F**) and resolved an immune/inflammatory program (MAZ[43]-NR2F6[44]-BHLHE40[45]) and a DNA damage/stress-response program (FOXO1[46]-E2F6[47]-ETS1[48]-MBD2[49]), the latter already implicated in AD[50–52], highlighting the potential relevance of stress-response pathways to microglial dysfunction.

MPRA reveal complementary biology across prioritization strategies

frVar were assigned to putative target genes (frGenes) using three complementary strategies: nearest-gene mapping, the microglial Activity-By-Contact (ABC) model, and microglial eQTLs (**Fig. 5B**). A total of 50 genes were linked to frVar by at least two of the three variant-to-gene approaches (highlighted at the bottom of the Miami plot in **Fig. 5A**). For each AD GWAS locus, we further annotated the number of functional variants detected and their corresponding frGenes (**Supplementary Fig. 6**).

Pathway enrichment using nearest-gene links was performed separately for GWAS-fine-mapped, eQTL-, and caQTL-prioritized variants, highlighting the context dependence of expression effects across microglial activation states, capturing homeostatic, inflammatory, and metabolic processes rather than converging on a single pathway (**Fig. 5C**). Whereas GWAS-fine-mapped variants were enriched for amyloid biology, immune/inflammatory responses, and cell-signaling pathways (**Fig. 5D, left**), aligning with established AD mechanisms, eQTL-prioritized variants mapped to a broader set of genes (**Fig. 5C, middle**), and caQTL-prioritized variants were enriched for neurodevelopment and structural organization (**Fig. 5D, right**). Together, these results underscore that GWAS-fine-mapped, eQTL and caQTL strategies highlight complementary and only partially overlapping aspects of AD biology.

Contrasting GO terms for genes linked to all variants (allGene) versus functionally validated frGenes in the AD GWAS category, revealed frGenes showed more specific enrichment for AD-related pathways and immune-synapse regulation (**Supplementary Fig. 7A**), whereas allGenes more broadly captured processes related to AD pathology, synaptic signaling, and neuroinflammation. frGenes were also preferentially expressed in “cluster 8” microglia from Sun et al.[53] single-cell data (postmortem AD; **Supplementary Fig. 7B**), a subpopulation associated with late-stage disease.

Finally, MPRA effects showed significant concordance with microglia eQTL effect sizes, but not whole-brain eQTLs, emphasizing the importance of cell-type-matched validation (**Supplementary Fig. 8**).

Functional validation of regulatory variants at the *TSPOAP1* locus

The power of MPRA is that it enables direct identification of functional regulatory variants, overcoming the limitations of *in silico* predictions. *TSPOAP1* represents a novel and understudied locus associated with AD. Statistical fine-mapping identified six putative causal variants within the credible set, with GWAS SNP rs2680700 showing the highest PIP, followed by rs2526377 (**Fig. 6, GWAS panel**). Another credible set contained a large number of variants with low PIPs, suggesting the presence of multiple independent signals that complicate fine-mapping interpretation (**Fig. 6, Fine-mapping panel**). This region overlaps multiple microglia-relevant annotations such as promoters, enhancers, and ATAC-seq peaks, highlighting its regulatory importance. (**Fig. 6, middle**). Colocalization between AD GWAS and microglial eQTL signals yielded a posterior probability (PP) of 0.55 for a shared causal signal and 0.15 for distinct but significant signals (**Supplementary Fig. 9**). In the MPRA, although rs2526377 was not top-ranked in GWAS or fine-mapping results, the alternative allele exhibited a clear repressive effect on transcriptional activity (**Fig. 6, MPRA panel**). Conversely, the microglial eQTL lead SNP rs116939255 also demonstrated regulatory activity, suggesting that the GWAS and eQTL associations at this locus may arise from distinct regulatory mechanisms.

Discussion

Here, we present the first systematic screen to examine the functional impact of AD-associated variants and eQTL variants in human microglia. By applying MPRA in both immortalized HMC3 cells and mature iMGL, we comprehensively characterized the regulatory potential of variants identified from GWAS, eQTL, and caQTL studies. frVar were significantly enriched in microglial regulatory regions compared with non-functional variants. They overlapped with TF binding sites, and TF-centered network analysis revealed two core regulatory modules. Overall, GWAS-prioritized variants were enriched for amyloid-related pathways, caQTL-prioritized variants emphasized neuronal structural development, and all categories converged on immune-related functions. Importantly, we demonstrate that MPRA can pinpoint putative functional variants in the novel AD loci such as *TSPOAP1*, and resolve multiple overlapping signals in a cell-type-specific manner. Together, this study establishes a scalable experimental framework for linking genetic risk variants to regulatory function and downstream pathways in disease-relevant human cell types, overcoming key limitations of purely computational approaches.

HMC3 is a widely used cellular model for human microglia[16–18], valued for its primary microglial origin and ease of maintenance. However, HMC3 cells are less reactive than typical immune cells[18] and show limited evidence of neuron-microglia interactions, a hallmark of microglial function and neuro-immune communication. Consistent with these limitations, distinct subsets of variants were identified in HMC3 compared with iMGLs, with only 21.5% of the HMC3 signals replicated in iMGLs, reflecting model-specific regulatory differences. Whereas iMGL showed a higher concentration of biological processes associated with microglia-neuron interaction, HMC3 functional genes were associated with signaling pathways and developmental terms more consistent with proliferation than reflecting terminal microglial functions. Using HMC3 as a model could therefore result in false positives from non-microglia-specific regulation, highlighting the importance of appropriate cell-type context for functional validation. We propose that iMGL models provide a more informative cellular context for dissecting regulatory mechanisms within microglia.

A key question in statistical genetics is whether integrating multiple variant prioritization methods better predicts functional activity[54–57]. Being comprised of ~25% AD GWAS-, ~50% eQTL-, and ~25% caQTL-prioritized variants, our MPRA revealed no clear advantage to any single prioritization strategy. Although variants with two or more annotations had a ~30% higher likelihood of being functional in MPRA compared to those with only a single annotation, only ~6% of variants carried more than one annotation, and just four variants were annotated by all three. The limited overlap aligns with our observation that GWAS-, eQTL-, and caQTL-prioritized functional variants captured distinct yet related biological processes in microglia. Together, these findings highlight that the prioritization strategies capture complementary aspects of microglial regulation with limited overlap. Relying solely on shared annotations may bias discovery toward specific pathways.

Despite the biological insights gained here, single-cell RNA-seq studies implicate distinct subtypes and spatial–temporal microglia heterogeneity in AD pathology[9], pointing to a major limitation of our study. It is probable that the cellular and environmental context is critical for MPRA, yet as an immune cell, microglia are prone to being reactive following introduction of foreign DNA. By design, MPRA can only evaluate variants outside their endogenous context, and so results may not accurately inform regulatory activity of distal gene targets. Likewise, the library only includes ~200 base pairs of sequence, potentially omitting key portions of larger regulatory regions. Together, these technical limitations potentially explain why most positive signals were enriched in transcription factor binding sites, being short-range sequence-based regulation. Altogether, MPRA does not directly infer causality, testing only if a variant regulates expression, but resolving neither the precise target genes nor their role in disease. These limitations highlight

290 the importance of integrating MPRA with complementary approaches like CRISPR[60,61]
291 perturbation and follow-up targeted validation.

292 Our results highlight how cellular models and variant prioritization methods can lead to
293 diverse regulatory results. Our findings offer a new perspective on the genetic influence on AD
294 and microglia biology. MPRA-informed functional maps can connect large-scale patient datasets
295 with the regulatory output and refine our understanding of AD risk prediction.

Methods

GWAS Data

We used the European Alzheimer's Disease Biobank (EADB) Stage I summary statistics for functional fine-mapping[2]. This GWAS analysis was based on 39,106 clinically diagnosed AD cases, 46,828 cases of proxy-AD and related dementia, and 401,577 controls. 21,101,114 variants passed quality control and variants below p=1x10-5 were selected. Conditional analyses (GCTA-COJO) revealed 75 loci and 82 lead variants. 33 of the loci were previously associated with AD and 42 loci were newly reported in this GWAS.

Fine-mapping Methods

For each lead signal/locus, we gathered all SNPs within 2-Mb windows (± 1 Mb flanking the lead GWAS SNP) and filtered out SNPs with a minor allele frequency (MAF) < 0.001. We focused on common variants in order to maximize the relevance of these results to a larger proportion of the AD population.

LD correlation matrices (in units of r) were acquired for each locus from the UK Biobank (UKB) reference panel, pre-calculated by Weissbrod et al.[23]. In addition, LD correlation matrices were calculated for each locus using the matched sample-level data from the EADB GWAS. Any SNPs that could not be identified within the UKBB + GWAS LD reference were removed from subsequent analyses.

Statistical fine-mapping was performed on each locus separately with FINEMAP[22] and SuSiE[62]. Functional fine-mapping was performed using PolyFun[63] + SuSiE and PolyFun + FINEMAP, both of which compute SNP-wise heritability-derived prior probabilities using an L2-regularized extension of stratified-linkage disequilibrium (LD) Score (S-LDSC) regression[64,65]. For PolyFun + SuSiE, we used the default UK Biobank baseline model composed of 187 binarized epigenomic and genic annotations. In all subsequent analyses presented here, SNPs that fall within the HLA region were excluded due to the particularly complex LD structure.

While the specifics of each fine-mapping model differ from tool-to-tool, they are united by several key features: 1) they are Bayesian models; 2) they provide the posterior inclusion probability (PP) that each SNP is causal, on a scale from 0 to 1, and 3) they provide credible sets (CS) of SNPs that have been identified as having a high PP of being causal, which we have set at a threshold of PP ≥ 0.95 for all tools. We only included tools that met the following criteria: 1) can take into account LD and 2) can operate using only summary statistics. For all fine-mapping models, we set the (maximum) number of causal SNPs to five.

ABC Model Bootstrap Analysis

We used Activity-by-Contact (ABC) model results from 131 different cell types[28]. Significant enhancers were determined by filtering out all enhancer-gene pairs that have an ABC score less than 0.02 in each cell-type (threshold determined by the original authors). We also use ABC model results that used microglia ATAC-seq and Hi-C data from the Kosoy et al.[7] paper, in order to determine cell-type specific enrichment in microglia compared to the 131 other cell types. A bootstrap test was performed to assess enrichment of AD fine-mapped variants in cell-type specific enhancers identified by ABC. We intersected variants with PP ≥ 0.1 in noncoding credible sets with ABC enhancers. For each cell-type, we calculate an enrichment ratio which is the number of fine-mapped variants that overlap ABC enhancers divided by the same number of randomly sampled common variants from the 1000 Genomes Project. We defined a cell-type as significantly enriched if the adjusted p-value (FDR) was <0.01.

Oligonucleotide Library Construction

341 The MPRA library preparation was adapted from the published lentiMPRA protocols[66]. Briefly, a
342 library of candidate regulatory sequences and adaptor sequences used for library construction
343 was synthesized as HiFi oligonucleotide library (~230 bp) from Agilent Technologies. These
344 oligonucleotides underwent two rounds of PCR for amplification and addition of random 15-bp
345 barcode sequences. The PCR products are cloned into a lentiviral vector upstream of a minimal
346 promoter and a eGFP reporter gene with NEBuilder HiFi DNA Assembly Master Mix (NEB
347 E2621X). Each candidate sequence was associated with an average of around 100 unique
348 barcodes to enable accurate quantification of reporter activity. The fragments were inserted into
349 the pLS-SceI vector (Addgene #137725) between SbfI and AgeI enzyme cutting site and then
350 transformed into NEB 10-beta Electrocompetent E. coli cells (NEB C3020K) via electroporation.
351 Bacterial stock containing a mixture of the library were grown overnight on Carbenicillin-positive
352 plates (Teknova L5010) and maxiprepped for plasmid collection (Invitrogen K210017). CRS and
353 barcode association are identified from the plasmid by Noveseq.

354 Microglia cellular culture

355 HMC3 is a cell-line purchased from American Type Culture Collection (ATCC CRL-3304) and
356 cultured in the Minimum Essential Medium media (Gibco 11095080) with 10% FBS. hiPSC-
357 derived microglia were differentiated from two control iPSC lines (#3182-XX and #2607-XY)[67].
358 The protocol was adapted from Abud[68] and McQuade[69] et al. The media from StemCell
359 Technologies were used to make hematopoietic stem and progenitor cells (HSPC) and then
360 differentiated into microglia. Briefly, hiPSC was grown in sparse aggregates with hematopoietic
361 media (STEMdiff™ Hematopoietic Kit #05310) for 12 days. Non-adherent cells CD43+ HSPC
362 were further plated with media (STEMdiff™ Microglia Differentiation Kit #100-0019)
363 supplemented with three key cytokines critical for microglia proliferation and maintenance. The
364 differentiating microglia were sub-cultured every 8 days and were collected on day 24 after plating
365 HPSCs into the microglial differentiation media. The final maturation step is achieved by the
366 addition of neuronal and astrocytic factors into the maturation media (STEMdiff™ Microglia
367 Maturation Kit #100-0020) and the matured iMGL can be collected after another 4 days of culture.
368 The iMGL were stained with markers to confirm its identity.

369 RNA and ATAC sequencing analysis

370 RNA-sequencing (RNA-seq) was performed in triplicate to characterize the transcriptional profile,
371 and ATAC-sequencing (ATAC-seq) was used to analyze the chromatin accessibility. Both RNA
372 and ATAC sequencing are done in the day 8 matured iMGL. Cells for RNAseq were harvested in
373 Triton X-100 (Sigma-Aldrich) and libraries were prepared using KAPA mRNA HyperPrep kit.
374 Paired-end sequencing reads (150bp) were generated with NovaSeq. Raw reads were aligned
375 to hg38 using STAR aligner (v2.7.11a)[70] and gene-level expression matrix were generated by
376 featureCounts[71] (Subread v2.0.3[72]) with GRCh38 genome and GENCODE[73] v40 annotation.
377 Genes with over 10 counts per million (CPM) in all sample were retained and the lowly expressed
378 genes were otherwise removed to ensure stable expression profile. Cells for ATAC were
379 harvested and freshly frozen in 100% FBS. DNA library was prepared with the Active Motif ATAC-
380 seq kit. Paired-end sequencing reads (150bp) were generative with NovaSeq. Adaptors in raw
381 reads were cut with cutadapt (v5.0)[74] and quality-checked with fastQC (v0.12.1)[75]. The
382 preprocessed fragments were aligned using Bowtie2 (v2.4.2)[76]. Mitochondria reads were
383 removed and BAM files were created, sorted, and indexed with SAMtools (v1.16)[77]. Peaks were
384 called with MACS2 (v2.2.9.1)[78] and mapping rate were checked with a customed python script.
385 Peaks were intersected between replicates and overlapping peaks were merged by BEDTools
386 (v2.30.0)[79].

387 MPRA lentivirus Transduction

The lentiviral library was packaged into second generation lentiviral particles by co-transfecting HEK293T cells with the library plasmid, along with helper plasmids encoding the necessary packaging proteins (psPAX2 and pMD2.G). The production of lentivirus containing our library is performed by Boston Children Hospital viral core. iMGLs were transducted on day 5 of the maturation session of the protocol and HMC3 is transducted when it is around 70% confluency. Virus-containing supernatants were harvested 48 hours post-transfection, filtered, and concentrated using lentiX concentrator. The MOI was calculated by the titration of lentivirus in different volumes of controlled virus (Addgene #137724) that were prepared in parallel of the MPRA virus. The genomic DNA was isolated and qPCR was used to detect integrated virus with background signal subtracted.

MPRA sequencing data analysis

The plasmid library was sequenced on a MiSeq instrument to link each candidate regulatory sequence (CRS) with its barcode. Pooled genomic DNA and RNA from the same culture after the MPRA experiment were sequenced on a NovaSeq. Sequencing data were processed using MPRAflow (version 2.0)[66] for initial read alignment, barcode counting, and CRS-barcode association. Only CRSs with at least 10 unique barcodes were retained. For each variant, incomplete reference-alternate pairs were removed. CRSs with a transcriptional rate below log2 RNA/DNA ≤ 0 were excluded, ensuring that only active constructs were analyzed.

We analyzed the resulting DNA and RNA count matrices using MPRAnalyze (version 1.24.0)[80], which employs a generalized linear model with a negative binomial distribution to jointly model RNA and DNA counts while accounting for library size and replicate depth. We estimated normalization factors and scaled both DNA and RNA libraries per replicate. Scrambled sequences were defined as controls and used to calibrate the null model in the scale mode. For allelic comparisons, we modeled RNA expression under reference vs alternative allele and testing for allelic effects using a likelihood-ratio test. The resulting $\log_2$ fold-change (logFC) represents the transcriptional difference of the alternate relative to the reference allele. Multiple testing correction was applied using the Benjamini–Hochberg method, and variants were called functional regulatory variants (frVar) if the allelic test reached FDR < 0.05 within each cell model.

Each MPRA was performed in three biological replicates for iMGLs derived from two donors (donor #3182-XX, n = 4) and for the HMC3 cell line (n = 3). Replicate correlations for iMGL across donors ranged from r = 0.88–0.98 for RNA, r = 0.85–0.96 for DNA, and r = 0.55–0.85 for RNA/DNA ratios. For HMC3, replicate correlations were r = 0.96 for RNA, r = 0.94–0.98 for DNA, and r = 0.80–0.90 for RNA/DNA ratios. Counts were aggregated across an average of ~100 barcodes per CRS prior to modeling. frVar were defined at the variant level, and the identified frVar were used as input for downstream analyses.

Variant functional annotation

The general ENCODE genomic annotations, including promoter, UTR, exon, intron, exon-intron boundaries information, were obtained from Bioconductor UCSC genomic features database (TxDb.Hsapiens.UCSC.hg38.knownGene)[81] and annotated with R package annotatr (v1.32.0)[82]. Both the entire library and functional subset was compared with randomly selected variant from the same chromosome with randomize_regions function. All results of enrichment are compared using Chi-squared test and adjusted for multiple testing using Benjamini-Hochberg correction.

Promoter and enhancer annotations for microglia were derived from primary human microglia[6]. Active promoters were identified by H3K27ac and H3K4me3 peaks overlapped; active enhancers were identified by H3K27ac peaks that were outside of H3K4me3 peaks. ATACseq peaks were obtained from matured iPSC-derived microglia. Enrichment differences between MPRA-positive and MPRA-negative variants were assessed using a Fisher's exact test.

GO-term enrichment analysis

All Gene Ontology (GO) annotations were obtained from org.Hs.eg.db (v3.19.1)[83], and enrichment analyses were performed using the clusterProfiler R package (v4.12.0)[84]. Nearest-gene mappings of functional variants were used as input, based on Ensembl IDs or gene symbols. Analyses were restricted to the Biological Process (BP) ontology. Genes expressed in iPSC-derived microglia or HMC3 cells served as the background set. Over-representation analysis was performed using a hypergeometric test, and p-values were adjusted using the Benjamini–Hochberg method (FDR < 0.05). Results were visualized with dot plots or bar plots.

Transcription factor motif disruption and GRN inference

TF binding motifs were obtained from the HOCOMOCO[85] human database. Variant–motif disruptions were evaluated using motifbreakR (v2.20.0)[86]. Only statistically significant motif hits were retained (filterp = TRUE), with a stringent p-value threshold of 1×10^{-4} to reduce false positives. The disruption score was calculated using the information content method, which accounts for the relative contribution of each nucleotide within the position weight matrix. To avoid bias from unequal nucleotide frequencies, a uniform background distribution of A, C, G, and T (0.25 each) was applied. Computations were parallelized using BiocParallel for efficiency.

For GRN construction, expression data from iMGL were filtered to include only genes expressed above 1 count per million (CPM) in at least 50% of samples, ensuring lowly expressed genes were excluded before normalization. Library sizes were adjusted using TMM normalization with edgeR, and log-transformed expression values were obtained as $\log_2$(CPM + prior count). GRNs were inferred using GENIE3 (v1.28.0)[87], which applies a random forest–based approach with 1,000 decision trees to capture stable regulatory interactions. At each tree split, the number of candidate regulators was set to the square root of the total number of regulators, balancing bias and variance. A fixed random seed was used to ensure reproducibility. The top 500 regulator–target edges ranked by importance score were visualized in Cytoscape. Gene identifiers were harmonized across datasets using biomaRt (v2.60.0)[88,89].

Locus plot

The composite locus plot is done with the plot_grid function in cowplot R package[90]. All plots were performed with GRCh38 data, if data was originated from GRCh37, all data is liftOver[91] from rtracklayer (v1.64.0)[92] before integration. GWAS summary statistics are from Bellengeuz et al[2], fine-mapping credible sets are from SuSiE fine-mapping result. Microglia promoter and enhancer data is from primary microglia activity-by-contact model built from Kosoy et al[7]. ATAC peak information is derived from ATAC sequencing data of iMGL. Gene models are from Ensembl (EnsDb.Hsapiens.v86, GRCh38)[93] and labeled by HGNC gene symbol.

References

1. Gatz, M. *et al.* Role of Genes and Environments for Explaining Alzheimer Disease. *Arch. Gen. Psychiatry* **63**, 168 (2006).
2. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
3. Schipper, M. & Posthuma, D. Demystifying non-coding GWAS variants: an overview of computational tools and methods. *Hum. Mol. Genet.* **31**, R73–R83 (2022).
4. Wang, Q. S. & Huang, H. Methods for statistical fine-mapping and their applications to auto-immune diseases. *Semin. Immunopathol.* **44**, 101–113 (2022).
5. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).
6. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
7. Kosoy, R. *et al.* Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat. Genet.* **54**, 1145–1154 (2022).
8. Askarova, A., Yaa, R. M., Marzi, S. J. & Nott, A. Genetic risk for neurodegenerative conditions is linked to disease-specific microglial pathways. *PLOS Genet.* **21**, e1011407 (2025).
9. Leng, F. & Edison, P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nat. Rev. Neurol.* **17**, 157–172 (2021).
10. Gao, C., Jiang, J., Tan, Y. & Chen, S. Microglia in neurodegenerative diseases: mechanism and potential therapeutic targets. *Signal Transduct. Target. Ther.* **8**, 359 (2023).
11. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* **12**, 1610 (2021).
12. Humphrey, J. *et al.* Long-read RNA sequencing atlas of human microglia isoforms elucidates disease-associated genetic regulation of splicing. *Nat. Genet.* **57**, 604–615 (2025).
13. Cao, X. *et al.* Integrative multi-omics QTL colocalization maps regulatory architecture in aging human brain. Preprint at https://doi.org/10.1101/2025.04.17.25326042 (2025).
14. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
15. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).
16. Dello Russo, C. *et al.* The human microglial HMC3 cell line: where do we stand? A systematic literature review. *J. Neuroinflammation* **15**, 259 (2018).
17. Karahan, H. *et al.* Deletion of *Abi3* gene locus exacerbates neuropathological features of Alzheimer's disease in a mouse model of Aβ amyloidosis. *Sci. Adv.* **7**, eabe3954 (2021).
18. Woolf, Z. *et al.* In vitro models of microglia: a comparative study. *Sci. Rep.* **15**, 15621 (2025).
19. Retallick-Townsley, K. G. *et al.* Dynamic neuro-immune regulation of psychiatric risk loci in human neurons. Preprint at https://doi.org/10.1101/2024.07.09.602755 (2024).
20. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
21. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genet.* **18**, e1010299 (2022).

515 22. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-
516 wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
517 23. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of
518 complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
519 24. Hooli, B. V. *et al.* The rare *TREM2* R47H variant exerts only a modest effect on Alzheimer
520 disease risk. *Neurology* **83**, 1353–1358 (2014).
521 25. Popescu, A. S. *et al.* Alzheimer's disease-associated R47H TREM2 increases, but wild-type
522 TREM2 decreases, microglial phagocytosis of synaptosomes and neuronal loss. *Glia* **71**,
523 974–990 (2023).
524 26. GERAD consortium *et al.* Increased expression of BIN1 mediates Alzheimer genetic risk by
525 modulating tau pathology. *Mol. Psychiatry* **18**, 1225–1234 (2013).
526 27. Chen, Z. *et al.* Context-dependent regulatory variants in Alzheimer's disease. Preprint at
527 https://doi.org/10.1101/2025.07.11.659973 (2025).
528 28. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**,
529 238–243 (2021).
530 29. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways
531 influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
532 30. Alzheimer Disease Genetics Consortium (ADGC), *et al.* Genetic meta-analysis of diagnosed
533 Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid
534 processing. *Nat. Genet.* **51**, 414–430 (2019).
535 31. Yuan, Y. *et al.* Differentially Expressed Heterogeneous Overdispersion Genes Testing for
536 Count Data*. Preprint at https://doi.org/10.1101/2023.02.21.529455 (2023).
537 32. Alzheimer's Disease Neuroimaging Initiative *et al.* Association of SPI1 Haplotypes with
538 Altered SPI1 Gene Expression and Alzheimer's Disease Risk. *J. Alzheimers Dis.* **86**, 1861–
539 1873 (2022).
540 33. Massopust, R. *et al.* KLF15 overexpression in myocytes fails to ameliorate ALS-related
541 pathology or extend the lifespan of SOD1G93A mice. *Neurobiol. Dis.* **162**, 105583 (2022).
542 34. García-Morales, V. *et al.* Sp1-regulated expression of p11 contributes to motor neuron
543 degeneration by membrane insertion of TASK1. *Nat. Commun.* **10**, 3784 (2019).
544 35. Li, D. *et al.* Distribution and functional significance of KLF15 in mouse cerebellum. *Mol.*
545 *Brain* **18**, 3 (2025).
546 36. An, G. *et al.* Loss of Bright/ARID3a Function Promotes Developmental Plasticity. *Stem*
547 *Cells* **28**, 1560–1567 (2010).
548 37. Medina-Martinez, O. *et al.* The transcription factor *Maz* is essential for normal eye
549 development. *Dis. Model. Mech.* dmm.044412 (2020) doi:10.1242/dmm.044412.
550 38. Hung, C.-Y. *et al.* Sp1 in Astrocyte Is Important for Neurite Outgrowth and Synaptogenesis.
551 *Mol. Neurobiol.* **57**, 261–277 (2020).
552 39. Cheng, Z.-Y., He, T.-T., Gao, X.-M., Zhao, Y. & Wang, J. ZBTB Transcription Factors: Key
553 Regulators of the Development, Differentiation and Effector Function of T Cells. *Front.*
554 *Immunol.* **12**, 713294 (2021).
555 40. Kim, B. *et al.* Effects of SPI1-mediated transcriptome remodeling on Alzheimer's disease-
556 related phenotypes in mouse models of Aβ amyloidosis. *Nat. Commun.* **15**, 3996 (2024).
557 41. Shao, J., Youngblood, H. & Yang, L. Targeting *SPI1* to mitigate amyloid-β pathology in
558 Alzheimer's disease. *J. Alzheimer's Dis.* **104**, 334–337 (2025).
559 42. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks
560 from Expression Data Using Tree-Based Methods. *PLoS ONE* **5**, e12776 (2010).

561 43. Xiao, T., Li, X. & Felsenfeld, G. The Myc-associated zinc finger protein epigenetically
562 controls expression of interferon-γ-stimulated genes by recruiting STAT1 to chromatin.
563 *Proc. Natl. Acad. Sci.* **121**, e2320938121 (2024).
564 44. Hermann-Kleiter, N. & Baier, G. Orphan nuclear receptor NR2F6 acts as an essential
565 gatekeeper of Th17 CD4+ T cell effector functions. *Cell Commun. Signal.* **12**, 38 (2014).
566 45. Cook, M. E., Jarjour, N. N., Lin, C.-C. & Edelson, B. T. Transcription Factor Bhlhe40 in
567 Immunity and Autoimmunity. *Trends Immunol.* **41**, 1023–1036 (2020).
568 46. Webb, A. E. & Brunet, A. FOXO transcription factors: key regulators of cellular quality
569 control. *Trends Biochem. Sci.* **39**, 159–169 (2014).
570 47. Bertoli, C., Herlihy, A. E., Pennycook, B. R., Kriston-Vizi, J. & De Bruin, R. A. M.
571 Sustained E2F-Dependent Transcription Is a Key Mechanism to Prevent Replication-Stress-
572 Induced DNA Damage. *Cell Rep.* **15**, 1412–1422 (2016).
573 48. Oikawa, T. & Yamada, T. Molecular biology of the Ets family of transcription factors. *Gene*
574 **303**, 11–34 (2003).
575 49. Zhang, Y. *et al.* The role of MBD2 in immune cell development, function, and autoimmune
576 diseases. *Cell Death Discov.* **11**, 280 (2025).
577 50. Iatrou, A., Clark, E. M. & Wang, Y. Nuclear dynamics and stress responses in Alzheimer's
578 disease. *Mol. Neurodegener.* **16**, 65 (2021).
579 51. Ranganathan, R., Sapozhnikov, G., Ni, W., Li, S. & Song, Y. Recent developments in the
580 role of DNA damage response and understanding its implications for new therapeutic
581 approaches in Alzheimer's disease. *Transl. Med. Aging* **7**, 52–65 (2023).
582 52. Delint-Ramirez, I. & Madabhushi, R. DNA damage and its links to neuronal aging and
583 degeneration. *Neuron* **113**, 7–28 (2025).
584 53. Sun, N. *et al.* Human microglial state dynamics in Alzheimer's disease progression. *Cell*
585 **186**, 4386-4403.e29 (2023).
586 54. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical
587 Fine-Mapping Studies. *PLoS Genet.* **10**, e1004722 (2014).
588 55. Anderson, D. & Lassmann, T. A phenotype centric benchmark of variant prioritisation tools.
589 *Npj Genomic Med.* **3**, 5 (2018).
590 56. Ming, J. *et al.* LSMM: a statistical approach to integrating functional annotations with
591 genome-wide association studies. *Bioinformatics* **34**, 2788–2796 (2018).
592 57. Gao, J., Espin-Garcia, O., Paterson, A. D. & Sun, L. Integrating variant functional annotation
593 scores have varied abilities to improve power of genome-wide association studies. *Sci. Rep.*
594 **12**, 10720 (2022).
595 58. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated
596 with Alzheimer's disease. *Nat. Commun.* **11**, 6129 (2020).
597 59. Wang, C. *et al.* A single-cell and spatial RNA-seq database for Alzheimer's disease
598 (ssREAD). *Nat. Commun.* **15**, 4710 (2024).
599 60. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–
600 2308 (2013).
601 61. Xu, Y. & Li, Z. CRISPR-Cas systems: Overview, innovations and applications in human
602 disease research and gene therapy. *Comput. Struct. Biotechnol. J.* **18**, 2401–2415 (2020).
603 62. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable
604 Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B*
605 *Stat. Methodol.* **82**, 1273–1300 (2020).

63. Weissbrod, O. *et al.* Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
64. Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
65. ReproGen Consortium *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
66. Gordon, M. G. *et al.* lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* **15**, 2387–2412 (2020).
67. Hoffman, G. E. *et al.* Transcriptional signatures of schizophrenia in hiPSC-derived NPCs and neurons are concordant with post-mortem adult brains. *Nat. Commun.* **8**, 2225 (2017).
68. Abud, E. M. *et al.* iPSC-Derived Human Microglia-like Cells to Study Neurological Diseases. *Neuron* **94**, 278-293.e9 (2017).
69. McQuade, A. & Blurton-Jones, M. Human Induced Pluripotent Stem Cell-Derived Microglia (hiPSC-Microglia). in *Induced Pluripotent Stem (iPS) Cells* (eds Nagy, A. & Turksen, K.) vol. 2454 473–482 (Springer US, New York, NY, 2021).
70. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
71. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
72. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108–e108 (2013).
73. Mudge, J. M. *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res.* **53**, D966–D975 (2025).
74. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
75. Andrews, S. FastQC: A Quality Control tool for High Throughput Sequence Data. (2020).
76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
77. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
78. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
79. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
80. Ashuach, T. *et al.* MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* **20**, 183 (2019).
81. Bioconductor Core Team, B. P. M. O. [Cre. TxDb.Hsapiens.UCSC.hg38.knownGene. Bioconductor https://doi.org/10.18129/B9.BIOC.TXDB.HSAPIENS.UCSC.HG38.KNOWNGENE (2017).
82. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).
83. Carlson, M. org.Hs.eg.db. Bioconductor https://doi.org/10.18129/B9.BIOC.ORG.HS.EG.DB (2017).
84. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

85. Vorontsov, I. E. *et al.* HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Res.* **52**, D154–D163 (2024).
86. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. *motifbreakR* : an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
87. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* **5**, e12776 (2010).
88. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
89. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
90. Claus Wilke *et al.* wilkelab/cowplot: 1.1.3. Zenodo https://doi.org/10.5281/ZENODO.10553544 (2024).
91. Bioconductor Package. liftOver. Bioconductor https://doi.org/10.18129/B9.BIOC.LIFTOVER.
92. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
93. Rainer, J. EnsDb.Hsapiens.v86. Bioconductor https://doi.org/10.18129/B9.BIOC.ENSDB.HSAPIENS.V86 (2017).

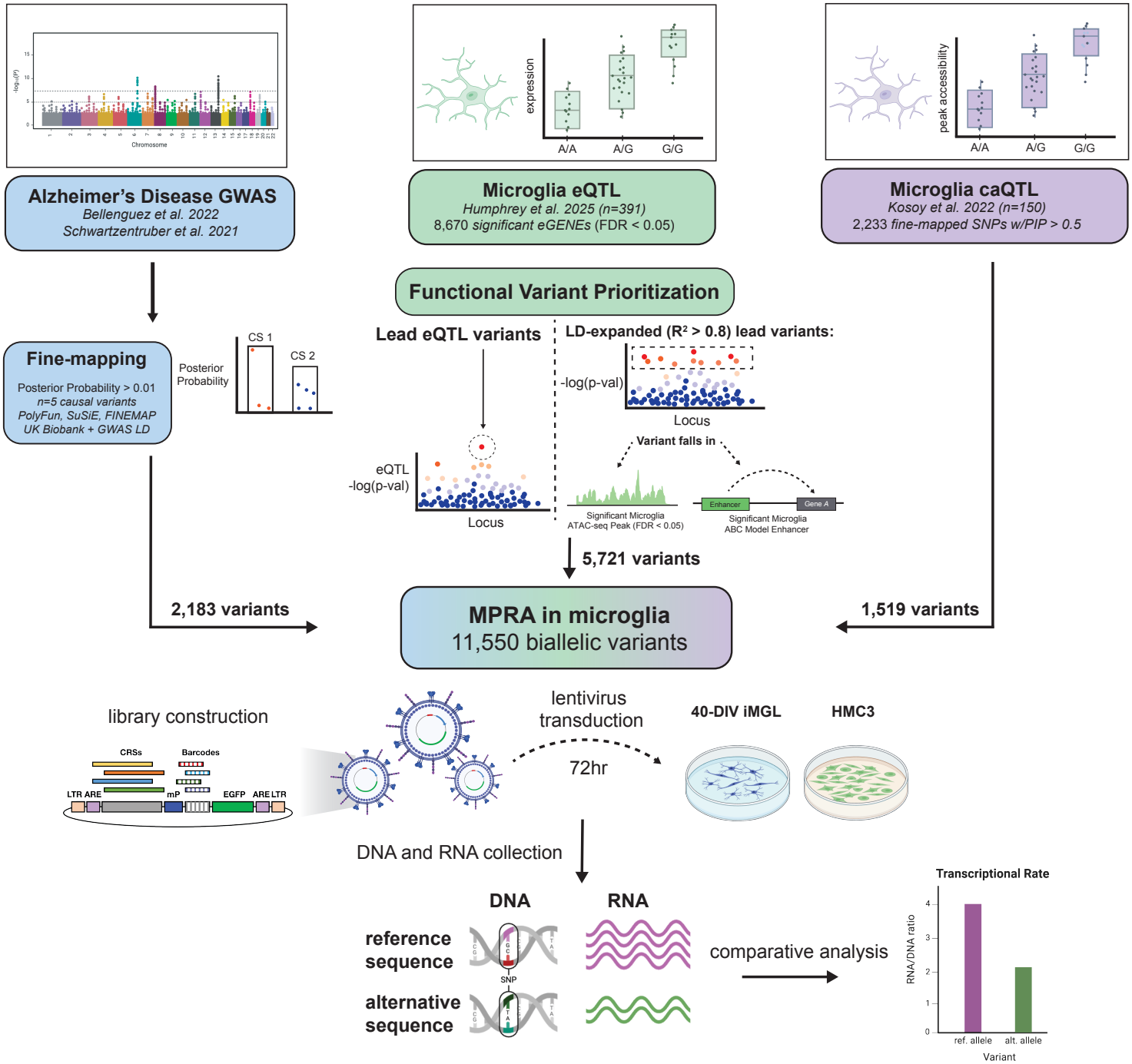

Figure 1. Overview of MPRA library design and experimental workflow.

Fine-mapped Alzheimer's disease (AD) GWAS variants, microglia eQTLs, and caQTLs were included in the MPRA library. Constructs consisted of the candidate regulatory sequence (CRS), a unique barcode, and a reporter gene. The library was packaged into lentivirus and introduced into microglial cell models (HMC3 and iPSC-derived microglia). DNA and RNA were extracted from the same cultures to quantify allele-specific transcriptional activity by comparing RNA/DNA ratios between reference and alternative alleles.

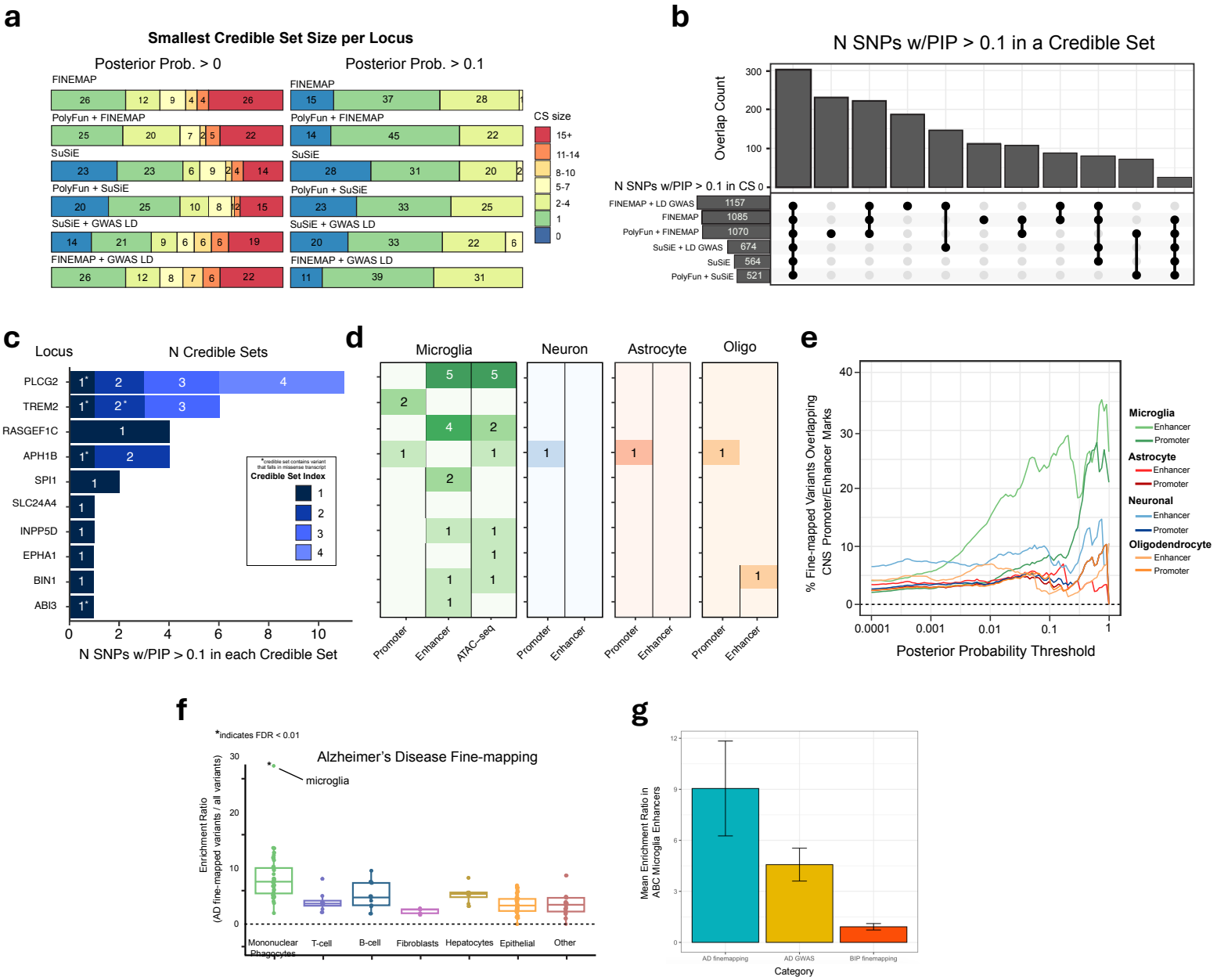

Figure 2. Summary of Fine-mapped Variants in Alzheimer's Disease (AD).

(a) Size of the smallest credible set per locus for all variants and variants with posterior probability > 0.1

(b) Upset plot showing the number of fine-mapped variants overlapped across methods

(c) Fine-mapped results across 10 subset AD GWAS loci

(d) Number of fine-mapped SNPs within each locus that overlap promoter and enhancer marks in microglia, neurons, astrocytes, and oligodendrocytes

(e) Fraction of fine-mapped variants above a given PP threshold that overlap promoter or enhancer marks across cell types in the central nervous system

(f) Enrichment ratio of AD fine-mapped variants within enhancers identified by the Activity-by-Contact (ABC) model from 132 cell types

(g) Enrichment ratio of AD fine-mapped variants, lead AD GWAS SNPs, and fine-mapped variants in bipolar disorder (BD) within ABC microglia enhancers

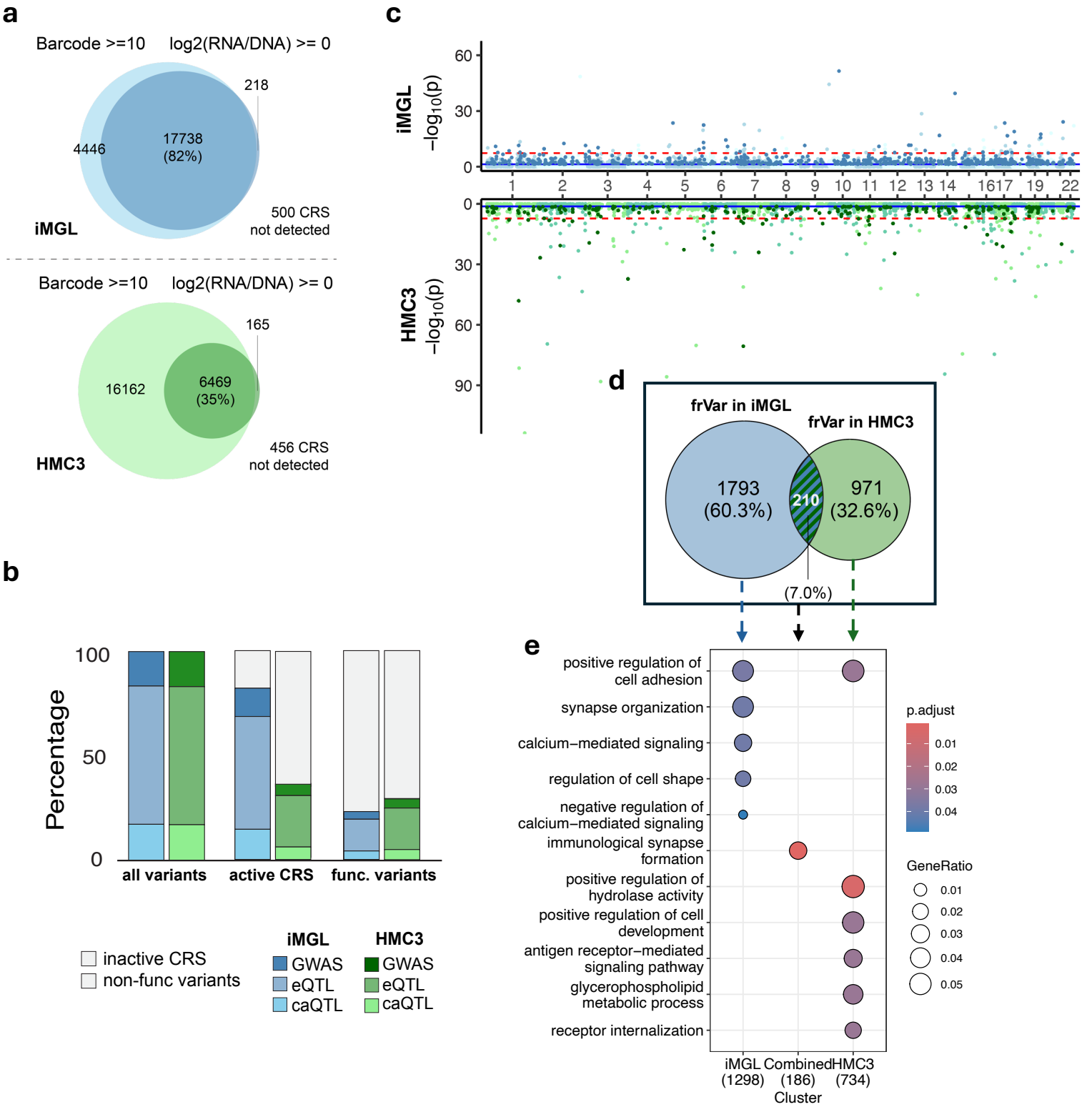

Figure 3. Characterization of Candidate Regulatory Sequences (CRS) and Functional Regulatory Variants Across Prioritization Methods and Cell Types

(a) Total number of CRS in the library that passed filtering for low barcode counts and low CRS activity.
(b) Percentage of CRS and frVar, categorized by their method of prioritization.
(c) Genome-wide distribution of functional regulatory variants (frVar).
(d) Overlap of frVar identified in induced microglia-like cells (iMGL) and HMC3 cells.
(e) GO-term enrichment analysis on the intersect and union of genes from two cell-lines.

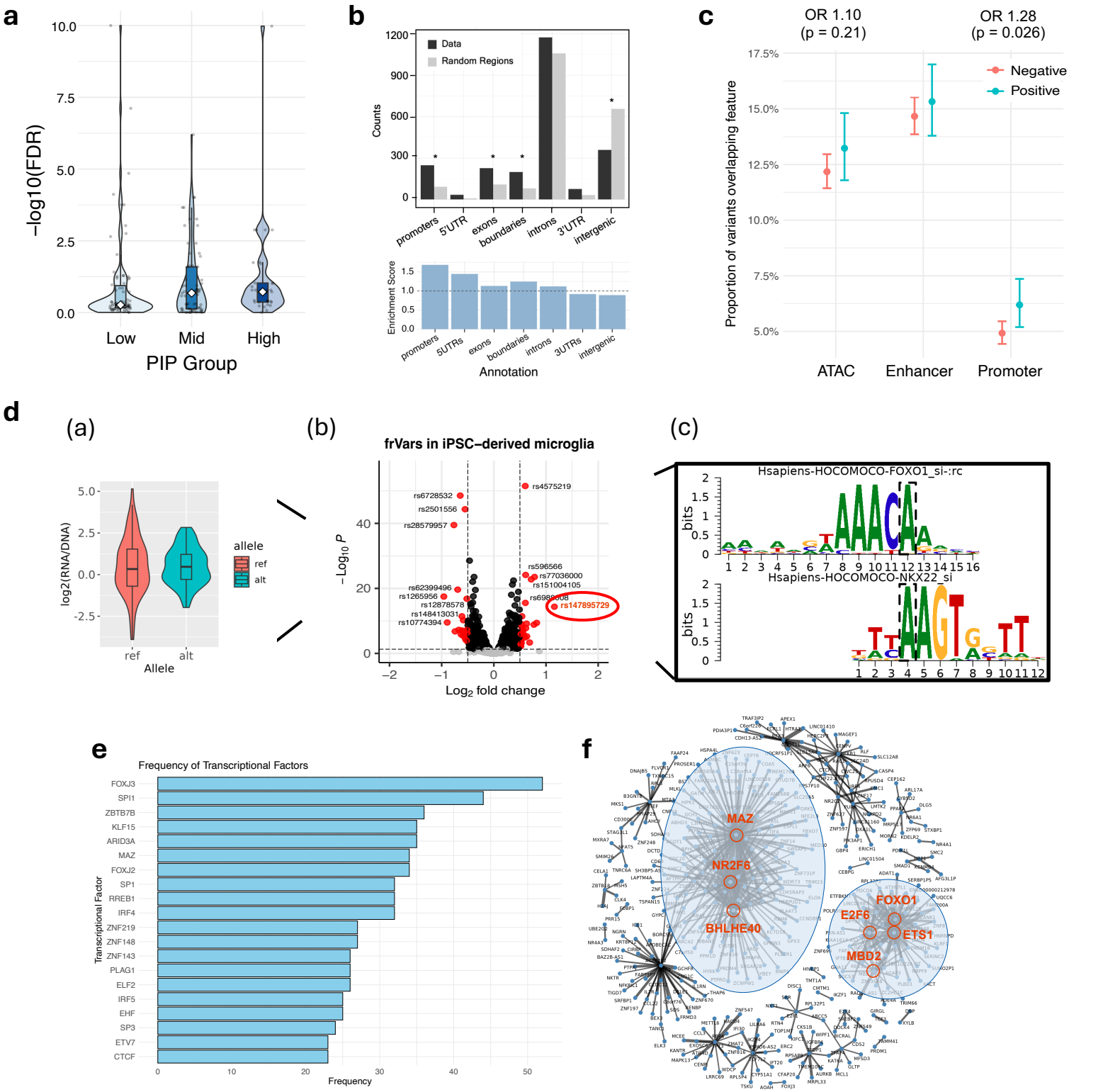

Figure 4. Function of Allele-Specific Regulatory Variants and Their Disease-Relevant Enrichment

(a) Distribution of significance by PIP group.

(b) Enrichment of functional regulatory variants (frVar) in genomic annotations and fold-change comparison between frVar and all tested variants (allVar).

(c) Enrichment of functional regulatory variants (frVar) in microglial annotation.

(d) a. Distribution of transcriptional activity for reference (REF) vs alternate (ALT) alleles at the example variant.
b. Volcano plot of allele-specific transcriptional activity across all tested variants.
c. Transcription factor (TF) binding motif disrupted by the example variant.

(e) TF binding site enrichment in CRSs containing frVar identified multiple Alzheimer's disease (AD)-associated TF as well as novel TF.

(f) TF-centered gene-regulatory network.

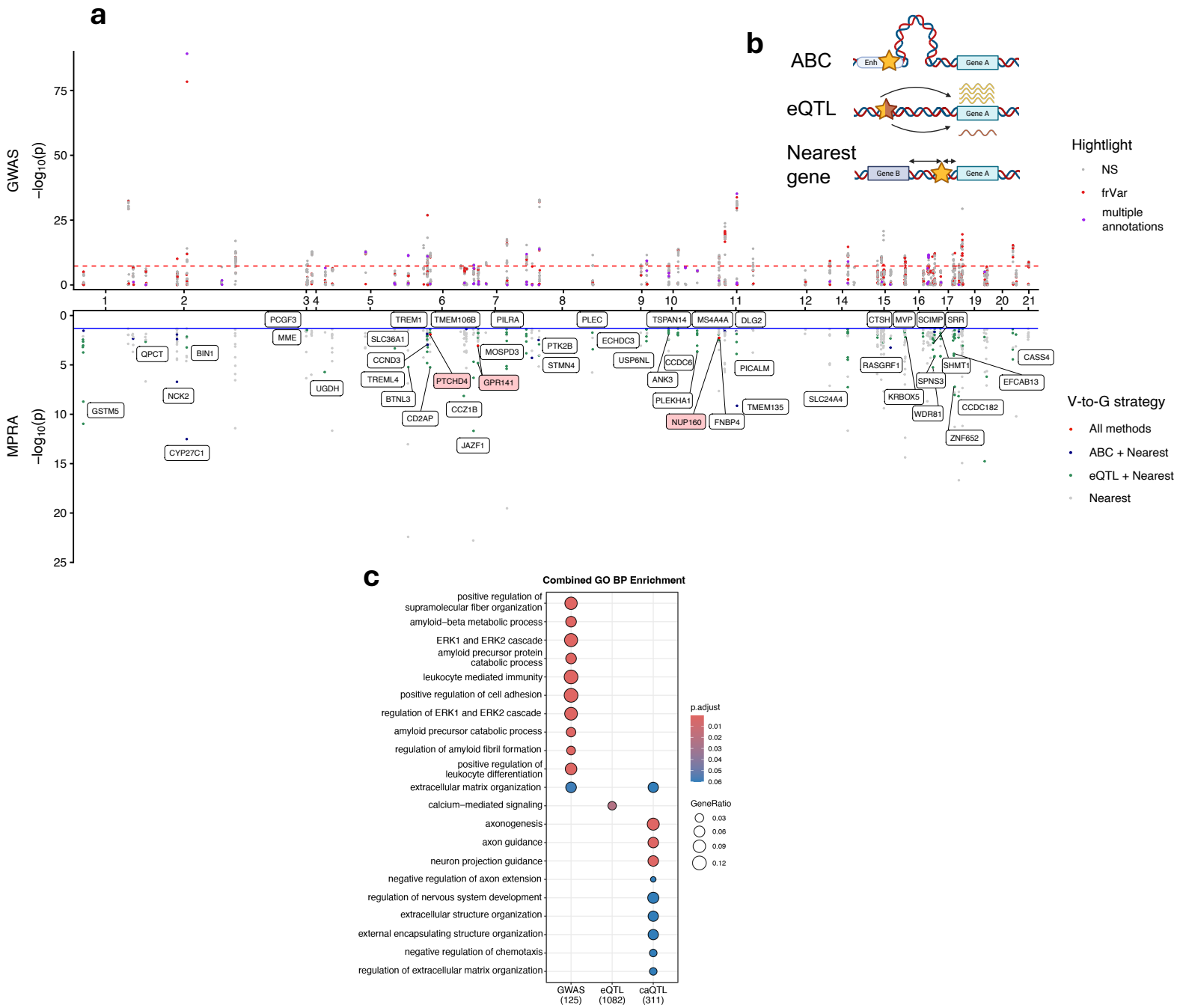

Figure 5. Variants prioritized by different methods reveal distinct biological regulation.

(a) Alzheimer's disease GWAS loci: the upper Manhattan plot showing GWAS results from Bellenguez et al.; red points indicate variants with significant regulatory activity identified by MPRA, and purple points highlight variants supported by multiple variant prioritization method. The lower Manhattan plot shows iMGL MPRA results; points are colored based on variant-to-gene (V-to-G) mapping strategies, and genes supported by multiple strategies are labeled.

(b) Schematic overview of the variant-to-gene mapping framework.

(c) Gene Ontology (GO) term enrichment analysis of genes prioritized by each method.

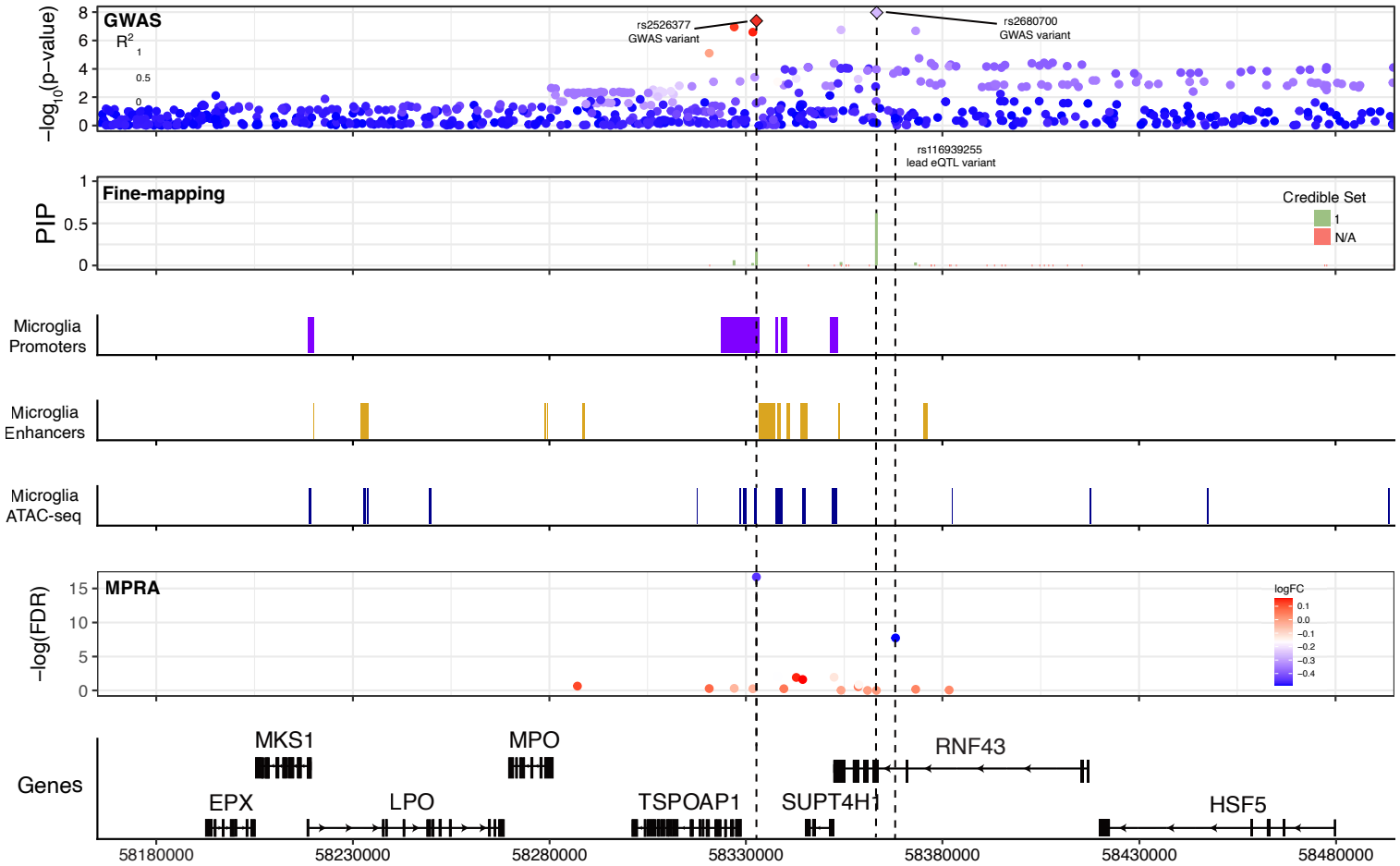

Figure 6. Novel AD loci example: TSPOAP1

This composite plot integrates multiple layers of genetic and functional evidence for one representative AD locus: (1) GWAS summary statistics from Bellenguez et al., colored by linkage disequilibrium with the lead SNP; (2) SuSiE fine-mapping results showing two credible sets; (3) chromatin accessibility and regulatory annotations, including primary microglia promoter and enhancer peaks and iMGL ATAC-seq peaks; (4) MPRA activity results for variants within the locus; and (5) gene models for all genes located in this region.