Title:

Functional regulatory variants implicate distinct transcriptional networks in dementia

Authors:

Yonatan A. Cooper ^{1,2,3}, Jessica E. Davis ⁴, Sriram Kosuri ⁴, Giovanni Coppola ^{3,5}, Daniel H. Geschwind ^{1,6,7,8,*}

Affiliations:

1. Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles; Los Angeles, CA, USA.
2. Medical Scientist Training Program, David Geffen School of Medicine, University of California Los Angeles; Los Angeles, California.
3. Center for Neurobehavioral Genetics, Jane and Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles; Los Angeles, CA, 90095, USA.
4. Department of Chemistry and Biochemistry, University of California, Los Angeles; Los Angeles, CA 90095, USA.
5. Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine University of California, Los Angeles; Los Angeles, CA, 90095, USA.
6. Program in Neurogenetics, Department of Neurology, University of California Los Angeles; Los Angeles, CA, USA.
7. Center for Autism Research and Treatment, Semel Institute, University of California Los Angeles; Los Angeles, CA, USA.
8. Institute of Precision Health, University of California Los Angeles; Los Angeles, CA, USA.

* Corresponding Author. Email: dhg@mednet.ucla.edu

Abstract:

Predicting functionality of noncoding variation is one of the major challenges in modern genetics. We employed massively parallel reporter assays to screen 5,706 variants from genome-wide association studies for both Alzheimer's disease (AD) and Progressive Supranuclear Palsy (PSP). We identified 320 functional regulatory polymorphisms (SigVars) comprising 27 of 34 unique tested loci, including multiple independent signals across the complex 17q21.31 region. We identify novel risk genes including *PLEKHMI* in PSP and *APOCI* in AD, and perform gene-editing to validate four distinct causal loci, confirming complement 4 (*C4A*) as a novel genetic risk factor for AD. Moreover, functional variants preferentially disrupt transcription factor binding sites that converge on enhancers with differential cell-type specific activity in PSP and AD, implicating a neuronal *SPI*-driven regulatory network in PSP pathogenesis. These analyses support a novel mechanism underlying noncoding genetic risk, whereby common genetic variants drive disease risk via their aggregate activity on specific transcriptional programs.

One Sentence Summary:

High-throughput functional analysis of GWAS loci reveals cell-type specific regulatory networks that mediate genetic risk for dementia.

Main Text:

Neurodegenerative disorders such as Alzheimer’s disease (AD) and Progressive Supranuclear Palsy (PSP) are a major and growing cause of morbidity and mortality worldwide, with AD alone expected to impact 135 million individuals by 2050 (1). Given the lack of any disease modifying therapeutics, there is a significant need for investigation of causal disease mechanisms. Sporadic AD and PSP, both known as tauopathies because of the pathological deposition of tau in the brains of affected individuals (2), are complex polygenic traits with heritability estimates of between 40-80% (3, 4). Over the last decade a number of genome-wide association studies (GWAS) have identified numerous susceptibility loci (5–12). However, most of these loci fall in noncoding regions of the genome and encompass numerous variants due to linkage disequilibrium (LD), which has hampered the identification of underlying regulatory variants and associated risk genes (13–15). This has posed a particular challenge for the interpretation of extended complex haplotypes harboring multiple independent association signals such as the H1 pan-neurodegenerative risk haplotype at 17q21.31 that includes the *MAPT* locus (12, 16, 17), and the AD risk locus at 19q13.32 that harbors *APOE* (18–20).

Although a number of statistical fine-mapping approaches have been developed to identify causal GWAS variants, these methods perform poorly on underpowered datasets or regions of extended LD (reviewed in (21)). Similarly, prioritization algorithms that score variant pathogenicity by leveraging features such as evolutionary conservation and chromatin annotations underperform in noncoding regions of the genome, or are nonspecific (22, 23). It is becoming increasingly recognized that functional methods are necessary to identify true causal variants within most loci, but the sheer numbers of variants challenge most experimental approaches. Massively parallel reporter assays (MPRA) provide a solution, enabling high-

throughput experimental characterization of the transcriptional-regulatory potential of noncoding DNA elements (24–27). In an MPRA, many regulatory elements are cloned into an expression vector harboring a reporter gene and a unique DNA barcode to create an expression library that is assayed via high-throughput sequencing (24). MPRA have prioritized functional common variation for lymphoblastoid eQTLs (28), red blood cell traits (29), cancer (30), adiposity (31), and osteoarthritis (32). However, they have not been systematically applied to neurodegeneration or any neurologic disorder.

In this study, we apply MPRA to characterize common genetic variation associated with two distinct neurodegenerative disorders that both share tau pathology, AD and PSP. We screen 5,706 unique variants comprising 34 genome-wide significant or suggestive loci derived from three GWASs (6, 7, 10), including a comprehensive assessment of disease-associated variants within the 17q21.31 and 19q13.32 regions. We identify 320 variants with significant allelic skew at a conservative false discovery rate and implicate novel risk genes including *PLEKHMI* in PSP and *APOC1* and *C4A* in AD. We find that SigVars preferentially disrupt transcription factor binding sites to impact specific transcriptional programs, and highlight a neuronal SP1 transcriptional network underlying PSP risk. Furthermore, using genome editing we validate rs13025717, rs636317, rs6064392, and rs9271171 as functional variants regulating *BINI*, *MS4A6A*, *CASS4*, and *HLA-DQA1/C4A* expression, respectively. Finally, we provide evidence that MPRA-determined variant prioritization is discordant with existing computational algorithms, underscoring the utility of high-throughput experimental datasets.

Results

MPRA to identify candidate regulatory GWAS variants

We conducted a staged analysis to identify regulatory variants underlying GWAS loci for two neurodegenerative tauopathies – Alzheimer’s disease (AD) and Progressive Supranuclear Palsy (PSP) – using massively parallel reporter assays (MPRA) (Fig. 1a). In stage 1 (MPRA 1), we identified all variants in linkage disequilibrium (LD; $r^2 > 0.8$) with the 1,090 genome-wide significant ($p < 5 \times 10^{-8}$) variants from an AD GWAS (6) and 3,626 genome-wide significant variants from a PSP GWAS (10). After filtering for bi-allelic noncoding variants, this resulted in a list of 5,223 variants encompassing 14 AD and 5 PSP GWAS loci. Both alleles of each variant were centered in 162 base pairs (bp) of genomic context, synthesized as 210 bp oligonucleotides, and cloned into a custom MPRA vector along with degenerate barcodes to create an expression library (Supplemental Fig. 1; Methods). In stage 2 (MPRA 2), we sought to replicate 326 variants screened in MPRA 1 to assess reproducibility, test the importance of oligo configuration on assay performance, and screen an additional 483 variants encompassing 11 new loci from 2 recent AD GWAS (6, 7) and 4 new suggestive loci for PSP (10) (Fig. 1a; Table 1; Methods).

The performance of MPRA to detect allelic skew depends upon high library transfection efficiency (33), necessitating the use of easy to transfect cell lines, which in published studies have included HEK293T, K562, and HepG2 cells, among others (28, 34). However, AD and PSP disease risk variants fall within open chromatin across several different neuronal and glial cell types, many of which are non-overlapping (7). Using available ENCODE DHS data (35), we found poor DNA accessibility overlap between divergent brain cell types, such as astrocytes and neural progenitors (mean Jaccard index = 0.14). Notably, HEK293T cells had the highest mean pairwise Jaccard index (0.22; Fig. 1b) compared with all brain cell types and tissues, which was

not driven by a larger number, or increased width of peaks in HEK293T cells. Moreover, we found that AD and PSP GWAS variants that fell within open chromatin in any brain cell type were also likely to fall within open chromatin in HEK293T cells (mean = 60%; Fig. 1c). These data indicated that HEK293T cells would provide an optimal model for such high-throughput screening in a single cell line, and they were chosen for the MPRA.

We performed MPRA (n = 6 biological replicates), obtaining activity measurements from at least 5 unique barcodes for both alleles for 4,732 of 5,223 variants (~91%), with a median library complexity of 40 barcodes/allele (Fig. S2a). Intra-library reproducibility was high; allele correlation between replicates had mean Pearson's $r = 0.95$ (Fig. 2b) and normalized barcode correlations were consistent with expectations ($r = 0.55$; Fig. S2c). Overall, we observed that ~19% of library elements were transcriptionally active in our assay (Fig. 2c), in concordance with previous estimates (28, 29).

We next assessed the functional genome annotations associated with active versus non-active elements. Using available ENCODE data for HEK293T cells, we found depletion of the H3K36me3 mark in active elements (Fisher's exact test; $\log_2$ OR = -0.54, FDR-adjusted $q = 0.01$). Conversely, active elements were highly enriched for DHS sites ($\log_2$ OR = 1.52, $q = 1 \times 10^{-10}$), which are indicative of accessible chromatin. Furthermore, H3K27ac and H3K4me3 marks, delineating active enhancers and promoters respectively, were likewise enriched in active elements (Fig. 2d; Fig. S2d). ChIP-seq peaks for specific Transcription Factors (TF) were similarly enriched in active elements (Fig. 2d; Fig. S2d). We also assessed for enrichment of specific TFBSs within active elements and identified significant enrichments of *SP/KLF*, *ETS*, and *AP-1* family members (Fig. 2e).

We applied a stringent statistical threshold to identify variants with significantly different transcriptional efficacy between alleles, termed *SigVars* (two-sided Mann-Whitney-U test, FDR $q < 0.01$, Benjamini-Hochberg method), identifying 267 *SigVars*. We next analyzed the second MPRA ($n = 6$ biological replicates), which maintained high quality with median library barcode complexity of 54 and high allele level (mean $r = 0.99$) and barcode level (mean $r = 0.84$) correlations between replicates (Fig. S3a-b). Assessments of reproducibility between the separate MPRA experiments also revealed that our assay was robust; the correlations between activity scores (Pearson's $r = 0.98$, $p < 2 \times 10^{-16}$) as well as effect sizes ($r = 0.94$, $p < 2 \times 10^{-16}$; Fig. 3a) for replicated variants were both high, and 152 of 186 (82%) re-tested *SigVars* were reproduced in the second MPRA (replication Bonferroni $p < 0.05$). Placing oligos in the reverse orientation completely abolished reporter activity as expected (Fig. S3c-d; Supplementary Text).

Given the high correspondence between both MPRA experiments, we combined them, obtaining activity measurements from 5,340 of 5,706 (93.6%; Supplemental Data) assayed variants and identifying 320 unique *SigVars* distributed across 17 chromosomes (6.0%; Fig. 3b). We identified *SigVars* in 27 of 34 (79%) tested GWAS loci with a median of 2 *SigVars* per locus (Fig. 3c; Table 1). As expected, effect sizes (alt/ref allele) were generally modest (Fig. 3d), with mean absolute *SigVar* $\log_2$ fold change of 0.53, consistent with prior work (29). *SigVars* were highly enriched within library elements that were transcriptionally active (Fig. 2c) in our screen ($OR = 6.2$; $p < 2 \times 10^{-16}$; Methods) and were also significantly enriched within DHSs within major brain cell types including astrocytes and neural progenitors, as well as monocytes, which are from the same lineage as microglia in the brain (Fig. 3e). However, when we separated *SigVars* derived from AD vs. PSP GWAS loci and identified those that were found to overlap enhancer marks from human brain neurons, microglia, astrocytes, and oligodendrocytes, we saw that the

cell types impacted by each disorder were distinct (Methods) (36). A plurality of AD SigVars fell within microglial enhancers. In contrast, a plurality of PSP SigVars fell within neuronal enhancers, in concordance with recent estimates of cell type specific enrichment in SNP-based heritability for these two disorders (37) (Fig. 3f). Interestingly, combined across both disorders, 55/78 (71%) of these variants overlapped with cell-type specific enhancer annotations (Fig. 3g).

Comparing computational predictions with empirically determined function

The use of computational methods for *a priori* identification of functional classes of variants is an ongoing area of active research and several predictive methods have been developed (38). We sought to determine the extent to which these methods recapitulate our empirical MPRA data and subsequent variant annotations, as well as previously published MPRA data in a different cell type from Tewhey *et al.* (28). We compared MPRA-determined effect sizes and allelic skew significance levels - derived from this study and the previously published MPRA dataset (28) - with regulatory scores from four widely used algorithms: CADD, CATO, GWAVA, and LINSIGHT (39–42). For both MPRA datasets, representing vastly different disorders and tissue types, we observed that these computational methods indeed were able to capture enrichment of regulatory signal when comparing the top vs bottom 5th percentile of variants as ranked by allelic skew p-values (Fig. S4a-b). However, if we applied a statistical threshold that would identify variants with functional effects (FDR-corrected thresholds of 0.01; Methods), overall regulatory predictions were highly discordant and not strongly predictive (max AUC = 0.55, 0.56; Fig. 4a-b). MPRA effect sizes also failed to correlate with algorithm predictions, with the exception of CATO (Fig. S4c-d). We also calculated pairwise overlap metrics and found that MPRA SigVar experimental outcomes had little overlap with

computational predictions (mean Cohen's Kappa = 0; Fig. 4c-d; Methods). Interestingly, prediction algorithms had little concordance among themselves; overall mean Kappa between all predicative algorithms was 0.11 for variants in our dataset and 0.14 for variants from the other published MPRA (28).

MPRA SigVars enrich for transcription factor binding site disruption

Because active MPRA elements enrich for TFBSs, we hypothesized that functional variants defined by our SigVars would be enriched for variants that disrupt TF-binding as a class. Therefore, we ran our variants through the SNPS2TFBS algorithm (43), which predicts TFBS disruption, finding that SigVars were enriched for variants that disrupt TFBSs (OR = 1.4, $p = 0.003$). We observed a similar magnitude of enrichment of TFBS-disrupting variants amongst SigVars from the published dataset from Tewhey *et al.* (28) (OR = 1.9, $p = 8.7 \times 10^{-8}$), and these findings were further reproducible using an alternative TFBS-scoring method (motifbreakR; Methods). Furthermore, the magnitude and directionality of predicted TFBS disruption correlated with MPRA effect sizes in both our dataset (Spearman's $\rho = 0.44$, $p = 2.3 \times 10^{-4}$) and the published dataset ($\rho = 0.52$, $p < 8.7 \times 10^{-9}$; Fig. 4e-f).

Given this enrichment, we asked whether TFBS disruption alone predicts MPRA allelic skew. To do so, we scored all variants from the previously published dataset (28) (assay performed in K562 cells) for TFBS disruption (Methods) and found that only a proportion of TFBS-disrupting variants were also MPRA SigVars (Positive Predictive Value = 0.14). Predictive performance somewhat improved after filtering for strong disruption of TFBSs for the top 200 K562-expressed TFs (PPD = 0.22; Methods), on par with the top-performing algorithm, CATO. These findings demonstrate the importance of accounting for cell-type specific *trans-*

factors for variant functional predictions, although predicted TFBS disruption only explains a subset of MPRA-defined functional variants. Interestingly, TFBS disruption was more predictive for some TFs over others. For example, disruption of *ELK1*, *ELF2*, and *GABPA* TFBSs (all ETS-family TFs), as well as SP-family TFBSs, were highly predictive of allelic skew captured by MPRA (Fig. S5), while disruption of other TF families was not.

Enrichment of functional risk variants within disease-specific transcriptional networks

We next assessed whether SigVars were enriched within binding sites for specific transcription factors. Importantly, we determined that TFBSs disrupted by risk variants differed by disease. In AD, *NR4A2* ($\log_2$ OR = 4.9, $p = 0.002$), *NR5A2* ($\log_2$ OR = 3.6, $p = 0.01$), *ATOH1* ($\log_2$ OR = 3.8, $p = 0.009$), *SP2* ($\log_2$ OR = 2.3, $p = 0.008$), and *SMAD*-family (*SMAD2,3,4* heterotrimer, $\log_2$ OR = 3.7, $p = 0.0002$) binding sites were enriched for disrupting risk variants (all FDR-adjusted $q < 0.05$). Interestingly, PSP showed a different pattern of TFBS enrichment, in which five of the six TFs predicted to be enriched for binding site disruption physically interact with the transcription factor SP1 (including SP1 itself; Fig. 4g). Importantly, while these TFs are highly expressed in HEK293T cells, we do not find a general relationship between relative TF abundance and MPRA allelic skew ($p > 0.05$; Spearman's correlation; Methods), indicating that expression levels do not explain our results. We also assessed enrichments at a lower statistical threshold (TFBS enrichment $q < 0.1$), finding a number of ETS-family TFs and SP1 interactors to be enriched within functional PSP risk variants (Fig. S6).

Further analysis indicated that the PSP-enriched TFs form a significant protein-protein interaction network with SP1 (permutation $p = 0.05$; Fig. 4g inset; Fig. S6; Methods), consistent with SP1's multimerization capabilities and its activity as a core component of a broad array of

gene regulatory complexes that regulate tissue specific gene expression (44). Interestingly, 11.1% of annotated PSP SigVars are predicted to participate specifically within this network (20% when including factors from the expanded network). Of note, the coding region for *SP1* itself falls within a suggestive PSP risk locus approaching genome-wide significance (locus combined p-value = 4.1×10^{-7}) (10). These data, including identification of five functional regulatory variants at the *SP1* locus, provide strong evidence for disruption of an SP1-based signaling network in PSP pathophysiology.

To explore this further, we next generated a two-layer directed network composed of these six PSP TFs and their likely targets, defined as genes within the *cis*-regulatory window of TFBS-disrupting SigVars (Fig 4h; Methods). Using available single-cell human brain gene expression data (45), we annotated all members of this network for their highest expressed cell type (Methods), and found that all but one of these genes were expressed most strongly in neurons (Fig. 4i), suggesting that this disrupted PSP-associated signaling network functions primarily within neurons, consistent with the observation of overall neuronal enrichment of heritable PSP risk variants from GWAS (46).

Refinement of SigVar annotations for high confidence predictions of causal variants

We next annotated the functional variants identified through our screen – representing likely causal variants – into those that fell within high confidence promoters or enhancers in neurons, astrocytes, microglia or oligodendrocytes (annotations (36)), as well as those variants that strongly disrupted predicted TFBSs (union of two algorithms (43, 47); Methods; Table S1). Of 320 SigVars, 233 (73%) had at least one additional functional annotation: 200 (63%) significantly altered TF-binding, 88 (28%) fell within a promoter or enhancer in at least one

brain cell type, and 55 (17%) were double annotated for TFBS disruption and an additional promoter/enhancer mark. These 55 SNPs, disrupting TFBSs within annotated promoters or enhancers, represent the highest confidence causal functional variants within ten distinct GWAS loci, including *BIN1*, *HLA-DRB1/5*, *PTK2B*, *CLU*, *ICQK/KNOX1*, 17q21.31/*MAPT*, 19q13.32/*APOE*, *ASAPI*, *SPI1/CELF*, and *CASS4*.

Systematic characterization of complex haplotypes 17q21.31 and 19q13.32

The chromosome 17q21.31 locus is noteworthy for harboring the tau-encoding *MAPT* gene within a common 900 kb inversion-polymorphism (*I6*), and is a major risk locus for PSP (H1 haplotype OR = 4-5) as well as AD, Parkinson's disease (PD), and Corticobasal Degeneration (CBD) (*I2*). 17q21.31 contains complex haplotypic sub-structural variation and extensive LD, hampering interrogation with traditional statistical genetics approaches. We leveraged the ability to functionally dissect this region, testing 3,482 variants within 17q21.31 in strong LD with lead SNPs from a PSP GWAS (*I0*), comprising approximately ~24% of the more than 14,000 common variants in the region. Of these, we identified a total of 194 SigVars, of which 111 were stringently replicated in both MPRA experiments. Of these, 20 variants were also double annotated for active chromatin features and TFBS disruption, making them very high confidence causal regulatory variants (Supplemental Table S1).

We next clustered these SigVars based on LD (BigLD algorithm)(48), which identified seven distinct LD clusters within four contiguous LD blocks (Fig. 5a), suggesting distinct loci within this region. The largest LD cluster includes 42 SigVars within *MAPT* itself (-5 kb upstream of TSS to 3' UTR) highlighting its striking regulatory complexity (Fig. 5b-c). Of these, 23 were replicated SigVars of which 13 are variants within annotated enhancers also predicted to

disrupt TFBSs (Fig. 5b). The *MAPT* promoter region overlaps a large CpG island and a number of repetitive transposon elements that may impact gene expression (49). This region has been previously characterized using serial deletion assays in a variety of cell types (50–52), which identified a core promoter beginning -300 bp from the TSS (49). In this study, the region -226/-63 (assay ID = 1447) was the 11th most transcriptionally active library element assayed overall, while -349/-186 (ID = 1446) had only modest expression (Supplemental Data), suggesting a more restricted core promoter starting at -186 upstream of the *MAPT* TSS (Fig. 5d). We also identified SigVars within the broader promoter region (-4364/+3292; Fig. 5c) including rs17770296, which falls in the distal promoter (-2612) and overlaps a MLT1I (ERVL-MALR family) transposable element. Another variant, rs76324150 (+1485), falls within neuronal H3K4me3 peaks and is predicted to disrupt binding of the TF *ZFX* (Fig. 5d).

SP1 is known to bind the *MAPT* core promoter and regulate Tau expression (53), and our analysis also demonstrates that a SP1 signaling network contributes to PSP pathogenesis (Fig. 4g-i). Unfortunately, four potentially interesting variants within the proximal promoter region (-144/+1485) dropped out of our assay likely due to the high GC content of the region (49) (Supplementary Text). Nevertheless, we identified four SigVars within the *MAPT* gene region predicted to disrupt binding of SP1 (Table S1), including rs76839282 which lies within H3K27ac peaks within the long regulatory intron 1 of *MAPT* (Fig. 5c).

We also identified SigVars within other LD blocks that are predicted to regulate independent risk genes. We highlight rs111392251, a high-confidence regulatory variant located in the promoter of *PLEKHMI* that is predicted to disrupt binding of *IRF*-family TFs (Fig. 5e-g). *PLEKHMI* regulates autophagosome-lysosome formation (54) and has been previously suggested as a PSP risk gene (9), but has yet to be extensively characterized in a disease context.

Other independent risk genes in distinct LD blocks implicated here include *MAP3K14* and *LRRC37A4P* (Table S2).

The *APOE* locus on 19q13.32 harbors the strongest common genetic association with late onset Alzheimer disease (LOAD), tagging the well-characterized *APOE4* risk haplotype (6, 19, 20). However, the extensive LD in the region coupled with the strength of the association signal has resulted in identification of hundreds of additional disease-associated variants (18, 19). Recent work involving transethnic scans and haplotype-aware conditional analyses have uncovered evidence for *APOE*- independent risk in the locus, implicating *PVRL2* and *APOC1* (18, 55), though others have argued that *APOE* coding variants mediate the entire association signal in the locus (19, 20, 56). We reasoned that identification of functional variants at 19q13.32 may help shed light on this complex regulatory architecture. We tested 640 variants in LD with the 538 genome-wide significant SNPs (6) at this locus, and identified 37 SigVars. Of these, at least 10 were whole blood or brain eQTLs for *PVRL2/NECTIN2* (GTEx)(57). These loci contained three intronic SigVars (rs34278513, rs412776, rs12972156), which were previously found to be significantly associated with LOAD when analysis was conditioned on *APOE4*-status (18). Our data identifies these variants as causal at this locus, and nominates their target gene as *PVRL2*. We also find that rs141622900, previously associated with cholesterol efflux capacity (58), is a SigVar residing in an active microglial enhancer directly downstream of *APOC1*, providing further support for *APOC1* as an AD risk gene. Finally, we identified an intergenic variant (rs2927437) within a robustly supported multi-tissue enhancer closest to the *BCL3* gene (59).

Validation of select MPRA SigVars

We next selected five high-confidence causal variants residing within clearly annotated regulatory regions for additional validation, two variants within well-characterized AD loci near *BIN1* and *MS4A6*, and three variants within less well-described loci near *HLA-DRB1/5*, *CASS4*, and *ECDHC3*. We used CRISPR-Cas9 genome editing to assay regulatory regions in their native genomic context by excising the enhancer elements containing these variants, then assaying downstream effects on gene expression using quantitative PCR (Methods). We identified the target genes for a given variant either when the variant was close to the gene within its cis-regulatory region, or those linked by chromatin interaction data based on Hi-C from a relevant cell-type (Methods).

We first assessed rs13025717, which is predicted to strongly disrupt binding of KLF4, resides ~20 kb from the TSS of the AD risk gene *BIN1*, and overlaps microglial and monocyte DHS as well as H3K27ac, H3K4me1, and H3K4me3 peaks (Fig. 6a-c). We used two pairs of gRNAs to excise a 240 or 374 bp segment containing rs13025717 in the SH-SY5Y cell line, and found a significant reduction in *BIN1* expression (Fig. 6d). Another intergenic variant, rs636317, lies within the Chr11:59923508 AD GWAS locus (*MS4A* gene region) (6) and overlaps H3K27ac and H3K4me1 peaks as well as DHS in microglia and monocytes (Fig. 6e). This variant is predicted to disrupt *CTCF* binding and is a highly significant MPRA SigVar (Fig. 6f-g). Analysis of existing Hi-C data from THP-1 cells (60) reveals rs636317 physically looping with upstream distal gene *MS4A6A*, which although not the closest gene to this variant, is also an eQTL (57) and the most highly-expressed gene at this locus in monocytes. As predicted, removing rs636317 in THP-1 cells, a macrophage-microglia related cell line (60) led to a significant reduction in expression of *MS4A6A* compared with both controls validating the function of this locus in a native context (Fig. 6h).

The *CASS4* gene, although identified within an AD-risk locus for close to a decade (6), remains relatively understudied. We identified functional variant rs6064392 within a microglial enhancer in an intergenic region upstream of *CASS4* (significant *ATF*-family TFBS disruption; Fig. 7a-d). The minor T-allele of rs6064392 is in tight LD ($r^2 = 0.91$) with the protective allele of rs6024870 (A; OR = 0.88; lead SNP)(7) and is predicted by MPRA to decrease downstream gene expression in agreement with eQTL data (Fig. 7c). We used a pair of gRNAs to excise a 382 bp region around rs6064392 in differentiated THP-1 cells, which significantly altered *CASS4* but not neighboring *RTF2* gene expression (Fig. 7e; Fig. S7a), supporting the functional impact of this variant on the *CASS4* gene.

Next, we explored the *HLA-DRB1/5* AD GWAS locus, which is a highly polymorphic region of extended LD with numerous AD-associated variants and potential risk genes. Our screen identified nine significant variants in the locus, the majority of which fell within the intergenic region between *HLA-DRB1* and *HLA-DQAI*. We chose to further validate rs9271171 as it fell within a large myeloid open-chromatin region and was a significant eQTL in whole blood for multiple genes (GTEx)(57) where the alternate (protective) allele decreased reporter expression. We used two pairs of gRNAs to excise part of the enhancer region containing the variant in differentiated THP-1 macrophages, which revealed a dramatic (2.5-fold; Fig. S7b) increase in expression of *HLA-DQAI*. Interestingly, unpublished chromatin conformation data from PsychENCODE indicated that rs9271171 might also physically interact with the distal gene, complement 4 (*C4A*), for which it is also an eQTL in whole blood (GTEx)(57). Indeed, we observed that excision of this variant significantly reduced *C4A* expression in THP-1 macrophages (Fig. 7g). Its regulation of both *HLA-DQAI* and *C4A*, but in opposite directions suggested that this region may be pleiotropic and regulate multiple downstream genes. Because

complement components are also expressed by reactive astrocytes, we performed the experiment again in human PSC-derived astrocyte cultures (Fig. S7c), confirming that excision of this variant reduced *C4A* gene expression in astrocytes as well. Lastly, the only variant that was tested, but not significant, rs7920721, was tested in THP-1 cells because of their similarity to microglia, but failed to impact gene expression for either *ECHDC3* or *USP6NL* (closest genes; Fig. S7d). However, we cannot rule out that this variant may function within a different cell type. In all, four of five variants significantly altered gene expression in the cis-regulatory region in the predicted gene of interest, validating these specific loci, and highlighting the general power of MPRA to identify functional variants.

Discussion

Predicting functionality of noncoding variation is one of the major challenges in modern genetics. In this study, we provide the first systematic characterization of common variants underlying disease risk for two distinct neurodegenerative disorders: Alzheimer's disease and Progressive Supranuclear Palsy. In doing so, we employ massively parallel reporter assays to screen 5,706 variants encompassing 34 unique loci identified across three genome-wide association studies (6, 7, 10), and obtain robust activity measures from 94% of tested elements. We find that active library elements are highly enriched for relevant chromatin and genomic features predictive of transcriptional regulation while confirming a high degree of inter- and intra-library reproducibility. We also utilize our two-stage study design to identify technical parameters impacting MPRA performance, adding to a growing literature on MPRA experimental design (61, 62) which will aid future studies (see Supplementary Text; Fig. S3c-d; Fig. S8). Most saliently, we identify 320 variants with significant transcriptional skew between alleles at a conservative false discovery rate ($q < 0.01$), thus delineating putative causal variants

at 27 of 34 tested genomic loci. We note that GWAS loci in regions harboring extensive LD and numerous risk genes are particularly difficult to interpret, and therefore require functional analyses. Here, we used MPRA to identify regulatory variants within two such regions, 17q21.31 and 19q13.32, risk factors for PSP and AD respectively, enabling identification of candidate regulatory variants. These data enabled us to identify regulatory variants within 29 and 17 distinct genes at 17q and 19q respectively (Table S3).

SNP-based heritability is known to be enriched within the regulatory regions within disease-relevant tissues (63). We demonstrate that functional regulatory variants further enrich within active, open chromatin, even relative to other GWAS variants within the LD block (Fig. 3e). Moreover, across two separate disorders, we find that a large proportion of functional variants fall within enhancers of distinct cell-types with a demonstrated relationship to disease, and that a large proportion of these enhancers are cell-type specific (71%; Fig. 3f-g).

We also determine that functional variants identified by MPRA are highly enriched for TFBS disruption, identifying transcriptional networks driving key aspects of disease risk (Fig. 4e-i; Fig. S6). For AD, this involves TFs including *NR4A2* and *SMAD*-family TFs, which have been described previously as acting within multiple cell-types in the brain and periphery to impact risk for AD (64, 65). For PSP, our analysis identified an enrichment of TFBS-disrupting functional variation in a significant protein interaction network with the transcription factor SP1. These TFs, as well as most of their predicted regulatory target genes are most highly expressed in neurons, consistent with cell-type specific aggregation of genetic risk in PSP. Moreover, we identified five SigVars in the PSP genome-wide suggestive locus harboring *SP1* (10), as well as four SP1 binding site-disrupting functional variants within the *MAPT* gene. These convergent findings provide strong evidence for disruption of SP1 signaling in neurons as a critical risk

factor for PSP. SP1 is known to regulate a broad array of cellular processes including chromatin remodeling, apoptosis, immune regulation, and response to oxidative stress in neurons (66). While SP1 network dysregulation has been identified in AD brain (67–69) it does not harbor genetic risk, and therefore is likely to play a reactive or secondary role. Overall, our data are consistent with PSP risk primarily impacting neurons, and astrocytes and oligodendrocytes secondarily, and AD risk in microglia and astrocytes, as has been reported (37, 46).

Our observations also suggest a refined model for understanding common genetic risk. Signal from intergenic GWAS loci are typically interpreted as deriving from causal regulatory variants that influence downstream expression of specific cognate risk genes, and thus are explainable through colocalization with eQTLs. Our results, implicating a TF network converging on SP1 in PSP, are consistent with a model whereby common genetic variants function in aggregate across multiple TFBSs to disrupt key cell-type specific transcriptional programs. We speculate that this genetic mechanism may particularly manifest itself only upon induction of the relevant transcriptional network, which may occur within a disease context. For example, transcriptomic and proteomic studies have previously identified induction of an SP1 transcriptional network in tauopathies (68, 69), which here we show may interact with common variation at relevant binding sites distributed across the genome to determine risk. It has been previously observed that GWAS loci and eQTLs exhibit limited genetic sharing, leading to a “missing” mechanism of action explaining a large proportion of noncoding loci (70). Our data show that polymorphisms altering binding of critical, disease-relevant transcriptional networks offer an additional explanation. That these transcriptional networks regulate a large number of cell-type enriched genes, provides a mechanism whereby genetic risk is expressed, not by

impacting a few core genes (71), but via polygenic cell type-specific regulatory effects on networks of genes (72).

A limitation of this study was the technical infeasibility of performing MPRA within human brain cell types, with the specific *trans*-regulatory environment of HEK293T cells likely influencing the generalizability of our screen (29). We addressed this by integrating cell-type specific regulatory annotations to prioritize variants likely to be functional within relevant brain cell types (Table S1). It should also be emphasized that this cell line showed the best overlap in active regulatory regions with multiple neuronal and glial cell types that contribute to disease risk, better than either a pure neuronal or glial cell type (Fig. 1b-c). We also identified and validated causal SNPs at four loci within brain-relevant cell lines, an important proof-of-principle for the external validity of our screen. In doing so, we provide strong evidence for rs13025717 and rs636317 as regulatory variants underlying the *BIN1* and *MS4A6A* loci respectively, as had been previously suggested (73, 74). Additionally, we identify rs6064392 as a likely causal variant within the *CASS4* locus, and our findings support previous reports that decreased *CASS4* expression may be protective in AD (75). Finally, we find evidence in two distinct cell types including iPSC-derived astrocytes that AD risk SNPs regulate *C4A*, the first time this gene has been implicated in common genetic risk for AD, further highlighting the importance of functional dissection of GWAS loci. We expect the combination of emerging eQTL and Hi-C data (57, 76) in conjunction with experimental CRISPR-mediated editing will enable further validation of other functional loci identified here, and their target genes. Additionally, some associated loci likely operate via mechanisms independent of direct transcriptional regulation, (e.g. splicing, post translational modification), requiring other approaches to delineate causal variants and their mechanisms. However, this study takes an

important step forward by providing the first large-scale functional annotation of these dementia risk loci.

References and Notes:

1. M. Prince, M. Guerchet, M. Prina, *The global impact of dementia 2013-2050* (Alzheimer's Disease International, 2013).
2. M. G. Spillantini, M. Goedert, Tau pathology and neurodegeneration. *Lancet Neurol.* **12**, 609–622 (2013).
3. M. Gatz, C. A. Reynolds, L. Fratiglioni, B. Johansson, J. A. Mortimer, S. Berg, A. Fiske, N. L. Pedersen, Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry.* **63**, 168–174 (2006).
4. S. L. Forrest, G. M. Halliday, H. McCann, A. B. McGeachie, C. V. McGinley, J. R. Hodges, O. Piguet, J. B. Kwok, M. G. Spillantini, J. J. Kril, Heritability in frontotemporal tauopathies. *Alzheimers Dement. Diagn. Assess. Dis. Monit.* **11**, 115–124 (2019).
5. D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. S. Pahwa, V. Moskvina, K. Dowzell, A. Williams, Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088 (2009).
6. J.-C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, G. Jun, A. L. DeStefano, J. C. Bis, G. W. Beecham, Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452 (2013).
7. B. W. Kunkle, B. Grenier-Boley, R. Sims, J. C. Bis, V. Damotte, A. C. Naj, A. Boland, M. Vronskaya, S. J. Van Der Lee, A. Amlie-Wolf, Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
8. I. E. Jansen, J. E. Savage, K. Watanabe, J. Bryois, D. M. Williams, S. Steinberg, J. Sealock, I. K. Karlsson, S. Hägg, L. Athanasiu, Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
9. G. U. Höglinger, N. M. Melhem, D. W. Dickson, P. M. Sleiman, L.-S. Wang, L. Klei, R. Rademakers, R. De Silva, I. Litvan, D. E. Riley, Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699 (2011).
10. J. A. Chen, Z. Chen, H. Won, A. Y. Huang, J. K. Lowe, K. Wojta, J. S. Yokoyama, G. Bensimon, P. N. Leigh, C. Payan, Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol. Neurodegener.* **13**, 1–11 (2018).

11. M. Y. Sanchez-Contreras, N. Kouri, C. N. Cook, D. J. Serie, M. G. Heckman, N. A. Finch, R. J. Caselli, R. J. Uitti, Z. K. Wszolek, N. Graff-Radford, Replication of progressive supranuclear palsy genome-wide association study identifies SLCO1A2 and DUSP10 as new susceptibility loci. *Mol. Neurodegener.* **13**, 37 (2018).
12. K. Bowles, D. A. Pugh, K. Farrell, N. Han, Y. Liu, S. A. Liang, L. Qian, J. Bendl, J. F. Fullard, A. E. Renton, 17q21. 31 sub-haplotypes underlying H1-associated risk for Parkinson's disease and progressive supranuclear palsy converge on altered glial regulation (2019).
13. L. D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
14. M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, Systematic localization of common disease-associated variation in regulatory DNA. *Science.* **337**, 1190–1195 (2012).
15. K. Watanabe, E. Taskesen, A. Van Bochoven, D. Posthuma, Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).
16. A. M. Pittman, A. J. Myers, J. Duckworth, L. Bryden, M. Hanson, P. Abou-Sleiman, N. W. Wood, J. Hardy, A. Lees, R. de Silva, The structure of the tau haplotype in controls and in progressive supranuclear palsy. *Hum. Mol. Genet.* **13**, 1267–1274 (2004).
17. M. Goedert, R. Jakes, Mutations causing neurodegenerative tauopathies. *Biochim. Biophys. Acta BBA-Mol. Basis Dis.* **1739**, 240–250 (2005).
18. X. Zhou, Y. Chen, K. Y. Mok, T. C. Kwok, V. C. Mok, Q. Guo, F. C. Ip, Y. Chen, N. Mullapudi, P. Giusti-Rodríguez, Non-coding variability at the APOE locus contributes to the Alzheimer's risk. *Nat. Commun.* **10**, 1–16 (2019).
19. G. Jun, B. N. Vardarajan, J. Buross, C.-E. Yu, M. V. Hawk, B. A. Dombroski, P. K. Crane, E. B. Larson, R. Mayeux, J. L. Haines, Comprehensive search for Alzheimer disease susceptibility loci in the APOE region. *Arch. Neurol.* **69**, 1270–1279 (2012).
20. A. C. Naj, G. Jun, G. W. Beecham, L.-S. Wang, B. N. Vardarajan, J. Buross, P. J. Gallins, J. D. Buxbaum, G. P. Jarvik, P. K. Crane, Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* **43**, 436–441 (2011).
21. D. J. Schaid, W. Chen, N. B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
22. L. Liu, M. D. Sanderford, R. Patel, P. Chandrashekar, G. Gibson, S. Kumar, Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 1–11 (2019).

23. K. Eilbeck, A. Quinlan, M. Yandell, Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
24. A. Melnikov, A. Murugan, X. Zhang, T. Tesileanu, L. Wang, P. Rogov, S. Feizi, A. Gnirke, C. G. Callan, J. B. Kinney, Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
25. C. D. Arnold, D. Gerlach, C. Stelzer, \Lukasz M. Boryń, M. Rath, A. Stark, Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. **339**, 1074–1077 (2013).
26. J. Ernst, A. Melnikov, X. Zhang, L. Wang, P. Rogov, T. S. Mikkelsen, M. Kellis, Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).
27. F. Inoue, N. Ahituv, Decoding enhancers using massively parallel reporter assays. *Genomics*. **106**, 159–164 (2015).
28. R. Tewhey, D. Kotliar, D. S. Park, B. Liu, S. Winnicki, S. K. Reilly, K. G. Andersen, T. S. Mikkelsen, E. S. Lander, S. F. Schaffner, Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*. **165**, 1519–1529 (2016).
29. J. C. Ulirsch, S. K. Nandakumar, L. Wang, F. C. Giani, X. Zhang, P. Rogov, A. Melnikov, P. McDonel, R. Do, T. S. Mikkelsen, Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*. **165**, 1530–1545 (2016).
30. S. Liu, Y. Liu, Q. Zhang, J. Wu, J. Liang, S. Yu, G.-H. Wei, K. P. White, X. Wang, Systematic identification of regulatory variants associated with cancer risk. *Genome Biol.* **18**, 1–14 (2017).
31. C. M. Vockley, C. Guo, W. H. Majoros, M. Nodzenski, D. M. Scholtens, M. G. Hayes, W. L. Lowe, T. E. Reddy, Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res.* **25**, 1206–1214 (2015).
32. J. C. Klein, A. Keith, S. J. Rice, C. Shepherd, V. Agarwal, J. Loughlin, J. Shendure, Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. *Nat. Commun.* **10**, 1–9 (2019).
33. A. Melnikov, X. Zhang, P. Rogov, L. Wang, T. S. Mikkelsen, Massively parallel reporter assays in cultured mammalian cells. *JoVE J. Vis. Exp.*, e51719 (2014).
34. M. Kircher, C. Xiong, B. Martin, M. Schubach, F. Inoue, R. J. Bell, J. F. Costello, J. Shendure, N. Ahituv, Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 1–15 (2019).

- 556 35. I. Dunham, E. Birney, B. R. Lajoie, A. Sanyal, X. Dong, M. Greven, X. Lin, J. Wang, T. W.
557 Whitfield, J. Zhuang, An integrated encyclopedia of DNA elements in the human genome
558 (2012).
- 559 36. A. Nott, I. R. Holtman, N. G. Coufal, J. C. Schlachetzki, M. Yu, R. Hu, C. Z. Han, M. Pena, J.
560 Xiao, Y. Wu, Brain cell type-specific enhancer-promoter interactome maps and disease-
561 risk association. *Science*. **366**, 1134–1139 (2019).
- 562 37. V. Swarup, T. S. Chang, D. M. Duong, E. B. Dammer, J. Dai, J. J. Lah, E. C. Johnson, N. T.
563 Seyfried, A. I. Levey, D. H. Geschwind, Identification of conserved proteomic networks in
564 neurodegenerative dementia. *Cell Rep*. **31**, 107807 (2020).
- 565 38. R. Ghosh, N. Oak, S. E. Plon, Evaluation of in silico algorithms for use with ACMG/AMP
566 clinical variant interpretation guidelines. *Genome Biol*. **18**, 225 (2017).
- 567 39. M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, J. Shendure, A general
568 framework for estimating the relative pathogenicity of human genetic variants. *Nat.*
569 *Genet*. **46**, 310–315 (2014).
- 570 40. M. T. Maurano, E. Haugen, R. Sandstrom, J. Vierstra, A. Shafer, R. Kaul, J. A.
571 Stamatoyannopoulos, Large-scale identification of sequence variants influencing human
572 transcription factor occupancy in vivo. *Nat. Genet*. **47**, 1393 (2015).
- 573 41. G. R. S. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Functional annotation of non-coding
574 sequence variants. *Nat. Methods*. **11**, 294–296 (2014).
- 575 42. Y.-F. Huang, B. Gulko, A. Siepel, Fast, scalable prediction of deleterious noncoding variants
576 from functional and population genomic data. *Nat. Genet*. **49**, 618–624 (2017).
- 577 43. S. Kumar, G. Ambrosini, P. Bucher, SNP2TFBS—a database of regulatory SNPs affecting
578 predicted transcription factor binding site affinity. *Nucleic Acids Res*. **45**, D139–D144
579 (2017).
- 580 44. P. Bouwman, S. Philipsen, Regulation of the activity of Sp1-related transcription factors.
581 *Mol. Cell. Endocrinol*. **195**, 27–38 (2002).
- 582 45. T. E. Bakken, N. L. Jorstad, Q. Hu, B. B. Lake, W. Tian, B. E. Kalmbach, M. Crow, R. D. Hodge,
583 F. M. Krienen, S. A. Sorensen, Evolution of cellular diversity in primary motor cortex of
584 human, marmoset monkey, and mouse. *bioRxiv* (2020).
- 585 46. V. Swarup, F. I. Hinz, J. E. Rexach, K. Noguchi, H. Toyoshiba, A. Oda, K. Hirai, A. Sarkar, N. T.
586 Seyfried, C. Cheng, Identification of evolutionarily conserved gene networks mediating
587 neurodegenerative dementia. *Nat. Med*. **25**, 152–164 (2019).

- 588 47. S. G. Coetzee, G. A. Coetzee, D. J. Hazelett, motifbreakR: an R/Bioconductor package for
589 predicting variant effects at transcription factor binding sites. *Bioinformatics*. **31**, 3847–
590 3849 (2015).
- 591 48. S. A. Kim, M. Brossard, D. Roshandel, A. D. Paterson, S. B. Bull, Y. J. Yoo, gpart: human
592 genome partitioning and visualization of high-density SNP data by identifying haplotype
593 blocks. *Bioinformatics*. **35**, 4419–4421 (2019).
- 594 49. M.-L. Caillet-Boudin, L. Buée, N. Sergeant, B. Lefebvre, Regulation of human MAPT gene
595 expression. *Mol. Neurodegener.* **10**, 1–14 (2015).
- 596 50. A. Andreadis, B. K. Wagner, J. A. Broderick, K. S. Kosik, A τ promoter region without
597 neuronal specificity. *J. Neurochem.* **66**, 2257–2263 (1996).
- 598 51. E. Sadot, A. Heicklen-Klein, J. Barg, P. Lazarovici, I. Ginzburg, *Identification of a tau promoter*
599 *region mediating tissue-specific-regulated expression in PC12 cells* (Academic Press, 1996).
- 600 52. B. Maloney, D. K. Lahiri, Structural and functional characterization of H2 haplotype MAPT
601 promoter: unique neurospecific domains and a hypoxia-inducible element would enhance
602 rationally targeted tauopathy research for Alzheimer’s disease. *Gene*. **501**, 63–78 (2012).
- 603 53. A. Heicklen-Klein, I. Ginzburg, Tau promoter confers neuronal specificity and binds Sp1 and
604 AP-2. *J. Neurochem.* **75**, 1408–1418 (2000).
- 605 54. D. G. McEwan, D. Popovic, A. Gubas, S. Terawaki, H. Suzuki, D. Stadel, F. P. Coxon, D. M. De
606 Stegmann, S. Bhogaraju, K. Maddi, PLEKHM1 regulates autophagosome-lysosome fusion
607 through HOPS complex and LC3/GABARAP proteins. *Mol. Cell*. **57**, 39–54 (2015).
- 608 55. A. M. Kulminski, I. Philipp, Y. Loika, L. He, I. Culminskaya, Haplotype architecture of the
609 Alzheimer’s risk in the APOE region via co-skewness. *Alzheimers Dement. Diagn. Assess.*
610 *Dis. Monit.* **12**, e12129 (2020).
- 611 56. G. R. Jun, J. Chung, J. Mez, R. Barber, G. W. Beecham, D. A. Bennett, J. D. Buxbaum, G. S.
612 Byrd, M. M. Carrasquillo, P. K. Crane, Transethnic genome-wide scan identifies novel
613 Alzheimer’s disease loci. *Alzheimers Dement.* **13**, 727–738 (2017).
- 614 57. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia,
615 N. Young, The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
- 616 58. C. Low-Kam, D. Rhinds, K. S. Lo, A. Barhdadi, M. Boulé, S. Alem, V. Pedneault-Gagnon, E.
617 Rhéaume, M.-P. Dubé, D. Busseuil, Variants at the APOE/C1/C2/C4 Locus Modulate
618 Cholesterol Efflux Capacity Independently of High-Density Lipoprotein Cholesterol. *J. Am.*
619 *Heart Assoc.* **7**, e009545 (2018).
- 620 59. J. Wang, X. Dai, L. D. Berry, J. D. Cogan, Q. Liu, Y. Shyr, HACER: an atlas of human active
621 enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106–D112 (2019).

60. D. H. Phanstiel, K. Van Bortle, D. Spacek, G. T. Hess, M. S. Shamim, I. Machol, M. I. Love, E. L. Aiden, M. C. Bassik, M. P. Snyder, Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. *Mol. Cell.* **67**, 1037–1048 (2017).
61. F. Inoue, M. Kircher, B. Martin, G. M. Cooper, D. M. Witten, M. T. McManus, N. Ahituv, J. Shendure, A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).
62. J. Klein, V. Agarwal, F. Inoue, A. Keith, B. Martin, M. Kircher, N. Ahituv, J. Shendure, A systematic evaluation of the design, orientation, and sequence context dependencies of massively parallel reporter assays. *Biorxiv*, 576405 (2019).
63. H. K. Finucane, B. Bulik-Sullivan, A. Gusev, G. Trynka, Y. Reshef, P.-R. Loh, V. Anttila, H. Xu, C. Zang, K. Farh, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
64. M. Moon, E. S. Jung, S. G. Jeon, M.-Y. Cha, Y. Jang, W. Kim, C. Lopes, I. Mook-Jung, K.-S. Kim, Nurr1 (NR4A2) regulates Alzheimer’s disease-related pathogenesis and cognitive function in the 5XFAD mouse model. *Aging Cell.* **18**, e12866 (2019).
65. T. Town, Y. Laouar, C. Pittenger, T. Mori, C. A. Szekely, J. Tan, R. S. Duman, R. A. Flavell, Blocking TGF- β –Smad2/3 innate immune signaling mitigates Alzheimer-like pathology. *Nat. Med.* **14**, 681–687 (2008).
66. N. Y. Tan, L. M. Khachigian, Sp1 phosphorylation and its regulation of gene transcription. *Mol. Cell. Biol.* **29**, 2483–2488 (2009).
67. B. A. Citron, J. S. Dennis, R. S. Zeitlin, V. Echeverria, Transcription factor Sp1 dysregulation in Alzheimer’s disease. *J. Neurosci. Res.* **86**, 2499–2504 (2008).
68. S. Canchi, B. Rao, D. Masliah, S. B. Rosenthal, R. Sasik, K. M. Fisch, P. L. De Jager, D. A. Bennett, R. A. Rissman, Integrating gene and protein expression reveals perturbed functional networks in Alzheimer’s disease. *Cell Rep.* **28**, 1103–1116 (2019).
69. V. P. Andreev, V. A. Petyuk, H. M. Brewer, Y. V. Karpievitch, F. Xie, J. Clarke, D. Camp, R. D. Smith, A. P. Lieberman, R. L. Albin, Label-free quantitative LC–MS proteomics of Alzheimer’s disease and normally aged human brains. *J. Proteome Res.* **11**, 3053–3067 (2012).
70. S. Chun, A. Casparino, N. A. Patsopoulos, D. C. Croteau-Chonka, B. A. Raby, P. L. De Jager, S. R. Sunyaev, C. Cotsapas, Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
71. E. A. Boyle, Y. I. Li, J. K. Pritchard, An expanded view of complex traits: from polygenic to omnigenic. *Cell.* **169**, 1177–1186 (2017).

- 657 72. N. R. Wray, C. Wijmenga, P. F. Sullivan, J. Yang, P. M. Visscher, Common disease is more
658 complex than implied by the core gene omnigenic model. *Cell*. **173**, 1573–1580 (2018).
- 659 73. G. Novikova, M. Kapoor, T. C. W. Julia, E. M. Abud, A. G. Efthymiou, H. Cheng, J. F. Fullard, J.
660 Bendl, P. Roussos, W. W. Poon, Integration of Alzheimer’s disease genetics and myeloid
661 cell genomics identifies novel causal variants, regulatory elements, genes and pathways.
662 *bioRxiv*, 694281 (2019).
- 663 74. M. R. Corces, A. Shcherbina, S. Kundu, M. J. Gloudemans, L. Frésard, J. M. Granja, B. H.
664 Louie, S. Shams, S. T. Bagdatli, M. R. Mumbach, Single-cell epigenomic identification of
665 inherited risk loci in Alzheimer’s and Parkinson’s disease. *bioRxiv* (2020).
- 666 75. S. J. Andrews, B. Fulton-Howard, A. Goate, Protective Variants in Alzheimer’s Disease. *Curr.*
667 *Genet. Med. Rep.* **7**, 1–12 (2019).
- 668 76. H. Won, L. de La Torre-Ubieta, J. L. Stein, N. N. Parikshak, J. Huang, C. K. Opland, M. J.
669 Gandal, G. J. Sutton, F. Hormozdiari, D. Lu, Chromosome conformation elucidates
670 regulatory relationships in developing human brain. *Nature*. **538**, 523–527 (2016).
- 671 77. D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T.
672 Doncheva, J. H. Morris, P. Bork, STRING v11: protein–protein association networks with
673 increased coverage, supporting functional discovery in genome-wide experimental
674 datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
- 675 78. A. T. Sørensen, Y. A. Cooper, M. V. Baratta, F.-J. Weng, Y. Zhang, K. Ramamoorthi, R. Froppf,
676 E. LaVerriere, J. Xue, A. Young, A robust activity marking system for exploring active
677 neuronal ensembles. *Elife*. **5**, e13918 (2016).
- 678 79. B. Bushnell, J. Rood, E. Singer, BBMerge—accurate paired shotgun read merging via overlap.
679 *PloS One*. **12** (2017).
- 680 80. J. Caballero, A. F. Smit, L. Hood, G. Glusman, Realistic artificial DNA sequences as negative
681 controls for computational genomics. *Nucleic Acids Res.* **42**, e99–e99 (2014).
- 682 81. R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L.
683 Hofacker, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
- 684 82. C. A. Sloan, E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K.
685 Narayanan, M. Ho, B. T. Lee, ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**,
686 D726–D732 (2016).
- 687 83. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic
688 features. *Bioinformatics*. **26**, 841–842 (2010).

84. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, \AAsa Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, Tissue-based map of the human proteome. *Science*. **347** (2015).
85. I. V. Kulakovskiy, I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res*. **46**, D252–D259 (2018).
86. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**, s13742-015 (2015).
87. C. Wang, M. E. Ward, R. Chen, K. Liu, T. E. Tracy, X. Chen, M. Xie, P. D. Sohn, C. Ludwig, A. Meyer-Franke, Scalable production of iPSC-derived human neurons to identify tau-lowering compounds by high-content screening. *Stem Cell Rep*. **9**, 1221–1233 (2017).
88. A. M. Patel, K. Wierda, L. Thorrez, M. van Putten, J. De Smedt, L. Ribeiro, T. Tricot, M. Gajjar, R. Duellen, P. Van Damme, Dystrophin deficiency leads to dysfunctional glutamate clearance in iPSC derived astrocytes. *Transl. Psychiatry*. **9**, 1–21 (2019).
89. T. D. Schmittgen, K. J. Livak, Analyzing real-time PCR data by the comparative C T method. *Nat. Protoc*. **3**, 1101 (2008).
90. M. C. Sachs, plotROC: A tool for plotting roc curves. *J. Stat. Softw*. **79** (2017).
91. F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, T. Manke, High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun*. **9**, 1–15 (2018).
92. C. Y. Lin, J. Lovén, P. B. Rahl, R. M. Paranal, C. B. Burge, J. E. Bradner, T. I. Lee, R. A. Young, Transcriptional amplification in tumor cells with elevated c-Myc. *Cell*. **151**, 56–67 (2012).

Acknowledgements:

We would like to thank Geschwind lab members Gayatri Nair and Samie Patel for their assistance with cell culture, and the UCLA TCGB and Flow Cytometry cores for their assistance and technical expertise.

Funding:

National Institute of Aging fellowship 1F30AG064832 (YAC)

720 UCLA-Caltech MSTP training grant T32-GM008042 (YAC)

721 National Institute of Neurological Disorders and Stroke grant 5UG3NS104095-04 (DHG, GC)

722 Rainwater Charitable Foundation award 20180629 (DHG, GC).

723 **Author contributions:**

724 Conceptualization: YAC, GC, DHG

725 Data Curation: YAC

726 Formal Analysis: YAC

727 Funding Acquisition: YAC, GC, DHG

728 Investigation: YAC

729 Methodology: YAC, JED, SK

730 Project Administration: YAC, GC, DHG

731 Resources: SK, GC, DHG

732 Software: YAC

733 Supervision: SK, GC, DHG

734 Validation: YAC

735 Visualization: YAC

736 Writing – original draft: YAC, DHG

737 Writing – review and editing: YAC, JED, SK, GC, DHG

738

739 **Competing interests:**

740 S.K. is an employee and holds equity in Octant, Inc. The authors have no additional competing

741 interests to declare.

742

Data and materials availability:

Raw and processed sequencing data for both MPRA experiments are available from GEO under accession GSE163855. To facilitate downstream analyses, preprocessed MPRA count data has also been uploaded to github: <https://github.com/ycooper27/Tauopathy-MPRA>. Processed and normalized MPRA data for both experimental stages are provided in Supplementary Table 1. Data underlying Figures 2c, 2d, 4a, 4c, 4g, 5a, 6d, 6h, 7e, 7g, and Supplementary Figures 4, 5, 6, and 7a are provided in Source Data.

Code availability:

Custom python scripts used for pre-processing MPRA data have been uploaded to github: <https://github.com/ycooper27/Tauopathy-MPRA>. Custom R code used for data analysis and visualization are available upon request.

Supplementary Materials:

Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S3
References (78-92)
Data S1-S3

Figures and Legends:

Fig. 1

a Project Workflow

Figure 1: (A) Project workflow: 1) 5,223 genome-wide significant variants and LD partners

769 encompassing 14 AD and 5 PSP GWAS loci were selected in MPRA 1. For MPRA 2, select
770 variants identified in MPRA 1 were replicated. An additional 483 variants from 11 AD and 4
771 PSP loci were also tested. 2) Both alleles of each variant were barcoded and cloned into an
772 expression library that was transfected into HEK293T cells. Allele expression was quantified by
773 next-generation sequencing of associated barcodes. 3) Strict quality control was performed to
774 confirm within-assay and between-experiment reproducibility. 4) Comparison of expression
775 between alleles enabled identification of variants with significant allele-specific transcriptional
776 skew (SigVars). 5) SigVars were used for transcription factor binding site-disruption analysis
777 and benchmarking of computational prediction algorithms. SigVars were further prioritized
778 using brain-specific genomic annotations and 6) five top variants were selected for validation
779 using CRISPR/Cas9 genome-editing in brain-relevant cell lines (THP-1, iPSC-astrocytes, SH-
780 SY5Y). **(B-C)** comparison of the chromatin landscape between divergent brain cell types. **(B)**
781 left – Heatmap quantifying overlap (Jaccard index) of DNase hotspots (ENCODE project)
782 between primary brain cell types and HEK293T cells. Right – boxplot displaying the mean
783 pairwise Jaccard index for each cell type. **(C)**, GWAS variants tested in this study were
784 overlapped with DNase hotspots from each cell type. The boxplot displays the mean pairwise
785 probability that a GWAS variant within open chromatin in a given cell type also overlaps open
786 chromatin in another cell type. Green arrows highlight HEK293T. Error bars = S.E.M.

Fig. 2

Figure 2: The MPRA exhibited strong technical performance with active MPRA elements enriched for functional chromatin annotations. **(A)** The bar charts (counts and percentages) show genomic annotations for 5,223 variants tested in MPRA 1. **(B)** MPRA 1 exhibits high reproducibility between biological replicates (mean $r = 0.95$; $n = 6$). Panels show inter-replicate, pairwise correlations of median $\log_2$ normalized barcode counts for all alleles passing filter ($n = 9,464$). Red line = regression line of best fit, Pearson's correlation, all $p < 2 \times 10^{-16}$. **(C)** Identification of active and repressed MPRA elements. MPRA library transcriptional activity was quantified by comparing the median normalized barcode count for each element against the median of the whole library ($n = 6$ replicates; one-sample Mann-Whitney-U test; two-sided, Bonferroni correction). Significantly (Bonferroni $p < 0.05$) increased (active) and decreased (repressed) library elements highlighted on the volcano plot in red and blue respectively. **(D-E)** Active elements are enriched for relevant chromatin features and TFBSs. **(D)** Enrichment $\log_2$ odds ratios (Fisher's exact test) of active and repressed elements within HEK293T ChIP-seq peaks for both histone and TF marks. Error bars = 95% CI, *** FDR-adjusted $q < 0.001$ (BH method), * $q < 0.05$. **(E)** Active elements enrich for predicted TF binding motifs, including *SP/KLF*, and *FOS/JUN* family TFs.

Fig. 3

Figure 3: Identification of variants with significant allelic skew (SigVars) from both MPRA stages. (A) Reproducibility of MPRA across experimental stages. 326 variants, including 186 SigVars identified from the preliminary MPRA (1) were re-tested in a follow-up MPRA (2). i) Reference allele transcriptional activity and ii) log₂ effect sizes (alt/ref allele) show strong correlation (Pearson's $r = 0.98, 0.94, p < 2 \times 10^{-16}$) between experiments. BC = barcode count. iii) 152 of 186 assessed SigVars from MPRA 1 were replicated in MPRA 2 at replication Bonferroni $p < 0.05$. (B) Manhattan plot of 5,340 unique variants successfully tested across both experiments. Red indicates SigVars at FDR-adjusted $q < 0.01$ (BH method). (C) Histogram of the number of SigVars identified per GWAS locus LD block (median = 2). (D) Volcano plot

814 shows $\log_2$ allelic skew effect sizes and $-\log_{10}$ p-values for 5,340 unique variants tested by
815 MPRA. SigVars for MPRA stage 1 (red) and 2 (blue) highlighted. blue line = median effect size.
816 (E) Enrichment $\log_2$ odds ratios (Fisher's exact test) of SigVars within DHSs of various brain
817 cell types (ENCODE project). Error bars = 95% CI , All FDR-adjusted $q < 0.001$. (F-G) SigVars
818 were separated into those derived from AD or PSP GWAS and annotated for overlap with
819 enhancer marks from sorted brain tissue (annotations (36)). Bar plot shows cell-type enhancer
820 annotation counts for overlapping SigVars separated by disease. A plurality of AD SigVars fell
821 within microglial enhancers, while PSP SigVars fell within neuronal enhancers. (G) most
822 SigVars overlap enhancers present in only one cell type.

Fig. 4

823

824 **Figure 4:** MPRA variant effect sizes correlate with TFBS disruption, but are poorly predicted by
825 computational algorithms. (A) ROC curves highlight the poor predictive performance (max AUC
826 = 0.55) of four algorithms used to score variant functionality (LINSIGHT, CADD, GWAVA,
827 and CATO) benchmarked against MPRA SigVars (defined at $q < 0.01$, BH method). (B) same as
828 (A) except using SigVars from a published MPRA dataset (28) (max AUC = 0.56). (C-D)
829 Heatmaps show pairwise Cohen's Kappa for MPRA SigVars and top-ranked variants scored by
830 each algorithm (Methods) for this study (C) and the published study (28) (D). (E-F) MPRA
831 variant effect sizes correlate with predicted TFBS-disruption scores for our dataset (E) and the
832 replication dataset (F). Red line = OLS regression, Spearman's rho. (G) Shows the $\log_2$
833 enrichments for significantly disrupted TFBSs (FDR $q < 0.05$, BH method), with SigVars from
834 AD and PSP analyzed separately (color = $-\log_{10}$ enrichment p-values, size = # disrupted TFBSs;
835 SNPS2TFBS (43)). Inset: i) protein-protein interaction network from the STRING (77) database
836 for significantly disrupted PSP-TFs. Line thickness = strength of evidence. ii) Empirical
837 distributions for expected PPI networkSP1 - connectivity generated by resampling (Methods).
838 Red lines = observed PSP-TF network edges. (H) Network of TFs disrupted in PSP and their
839 target genes (genes within 10 kb of disrupted TFBSs). (I) Bar plot counts top-expressing brain
840 cell-type for each gene in (H) colored by network layer (cyan = TF, red = target gene).

Figure 5: Systematic dissection of functional variation at 17q21.31. (A) Top: LD plot for common variants in 17q21.31 (1000 Genomes, CEU). MPRA SigVars were clustered by LD (CLQD clustering (48); Methods). Bottom: SigVars plotted by position across 17q21.31 and

845 MPRA significance ($-\log_{10}$ p-values; colors = cluster annotations, black variants are unclustered).
846 Most variants fall within LD-cluster 4 (purple) centering on *MAPT*. **(B)** Annotation breakdown
847 for *MAPT* SigVars. ChIP-seq from (36), TFBS-X = union of predicted TFBS-disruption from
848 two algorithms (43, 47)). **(C)** Chromatin annotations and MPRA SigVars across *MAPT*.
849 Genomic tracks (1-8) show H3K27ac (enhancer) and H3K4me3 (promoter) ChIP-seq peaks for
850 microglia (red), neurons (green), astrocytes (blue), and oligodendrocytes (orange) from sorted
851 human brain tissue (36), CpG islands (track 9), and all tested variants plotted by significance ($-\log_{10}$ p-value, track 10). SigVars (FDR $q < 0.01$, BH method) highlighted in red. Track 11 shows
852 the *MAPT* gene model. Exons: Untranslated = blue, constitutive = white, alternative neuronal =
853 green, rarely expressed in brain = yellow. **(D)** Functional annotations of the *MAPT* promoter ($-4364/+3292$, numbering relative to TSS, features from (52)) with SigVars shown (red dots). Red
854 = repressor regions. Bottom: MPRA relative reporter activity for two elements. **(E-G)** highlight
855 rs111392251, a SigVar in the promoter (± 3 kb) of *PLEKHMI*. **(E)** genomic tracks (1-4) for
856 H3K4me3 peaks (as in **B**). **(F)** The alternate allele of rs111392251 is predicted to significantly
857 disrupt the TFBS for *IRF1-3* TFs. **(G)** violin plots show normalized barcode distributions for
858 each allele (MPRA FDR-adjusted q-values shown).
861

Fig. 6

Figure 6: Validation of rs13025707 and rs636317 as regulatory variants underlying *BIN1* and *MS4A6* GWAS loci. **(A-D)** convergent evidence for rs13025707 as a regulatory variant. **(A)** Genomic tracks (1-4) at the *BIN1* locus show rs13025707 falling within H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by $-\log_{10}$ p-value (track 5). SigVars (FDR $q < 0.01$, BH method) shown in red. **(B)** the alternate allele of rs13025707 is predicted to disrupt binding of *KLF4*. **(C)** violin plots show normalized barcode distributions for each allele across both MPRA stages. FDR-adjusted p-values displayed. **(D)** CRISPR-mediated deletion of small genomic regions containing rs13025707 in SH-SY5Y cells significantly reduces *BIN1* expression compared with gRNA-scramble controls ($n=3/\text{group}$, combined $n=6/\text{condition}$; $t(10) = -6.0$; $p = 0.0001$; two-tailed Student's t-test). Error bars = S.E.M. **(E-H)** convergent evidence for rs636317 as a regulatory variant. **(E)** Genomic tracks (1-5) at the *MS4A6* locus show rs636317 falling within CTCF, H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by $-\log_{10}$ p-value (track 6). SigVars shown in red. Monocyte Hi-C data (60) (track 6) links rs636317 with *MS4A6A*. **(F)** rs636317 is predicted to disrupt binding of *CTCF*. **(G)** violin plots show normalized barcode distributions for each allele across both MPRA stages. FDR-adjusted p-values shown. **(H)** CRISPR-mediated deletion of a genomic region containing rs636317 in differentiated THP-1 monocytes significantly reduces *MS4A6A* expression ($n=4/\text{group}$; $t(10) = -2.3$; $p = 0.03$; two-tailed Student's t-test). Error bars = S.E.M.

Fig. 7

Figure 7: Validation of rs6064392 and rs9271171 as regulatory variants underlying *CASS4* and *HLA-DRB1/5* GWAS loci. **(A-E)** convergent evidence for rs6064392 as a regulatory variant. **(A)** Genomic tracks (1-4) at the *CASS4* locus show rs6064392 falling within H3K27ac, H3K4me1, and DHS peaks (CD14+ monocytes; ENCODE) upstream of the *CASS4* TSS. All variants tested by MPRA in the locus plotted by $-\log_{10}$ p-value (track 5). SigVars (FDR $q < 0.01$, BH method) shown in red. **(B)** rs6064392 is predicted to disrupt binding of *ATF6*. **(C)** Whole blood eQTL (GTEx) for rs6064392 and *CASS4*. **(D)** violin plot shows normalized barcode distributions for each allele (MPRA 2). FDR-adjusted p-values displayed (two-tailed Mann-Whitney-U test). **(E)** CRISPR-mediated deletion of a small genomic region containing rs6064392 in differentiated THP-1 cells significantly reduces *CASS4*, but not *RTF2* (Supplemental Fig 6a) expression compared with gRNA-scramble controls ($n=4/\text{group}$; $t(6) = -4.3$; $p = 0.005$; two-tailed Student's t-test). Error bars = S.E.M. **(F)** Genomic tracks (1-4) at the *HLA-DRB1/5* locus show rs9271171 falling within H3K27ac, H3K4me1, H3K4me3, and DHS peaks (CD14+ monocytes; ENCODE). All variants tested by MPRA in the locus plotted by $-\log_{10}$ p-value (track 6). SigVars shown in red. **(G)** CRISPR-mediated deletion of genomic regions containing rs9271171 in differentiated THP-1 monocytes significantly reduces *C4A* expression relative to controls ($n=4$ group, $n=8/\text{condition}$; $t(14) = 3.6$; $p = 0.003$; two-tailed Student's t-test). Error bars = S.E.M.

Table 1 – AD and PSP GWAS Loci

Chr	Pos/Cytoband	Lead SNP	Annotated Gene	GWAS [#]	MPRA Stage	# SNPs Tested	SigVars
1	85603051	rs114573015	<i>WDR63</i>	PSP	2	4	0
1	180993146	rs57113693	<i>STX6</i>	PSP	1	44	2
1	207692049	rs6656401	<i>CRI</i>	AD 1	1	22	2
1	221995092	rs12125383	<i>DUSP10</i>	PSP	2	51	3
2	127892810	rs6733839	<i>BIN1</i>	AD 1	1	55	7
2	234068476	rs35349669	<i>INPP5D</i>	AD 1	2	45	2
3	39510287	rs10675541	<i>MOBP</i>	PSP	1	80	2

6	32578530	rs9271192	<i>HLA-DRB5- HLA-DRB1</i>	AD 1	1	445	13
6	45499614	rs35740963	<i>RUNX2</i>	PSP	1	65	1
6	47487762	rs10948363	<i>CD2AP</i>	AD 1	1	71	3
7	100004446; 100091795	rs1476679; rs12539172	<i>ZCWPW1; NYAP1</i>	AD 1&2	2	7	0
7	143110762	rs11771145	<i>EPHA1</i>	AD 1	1	29	3
8	27195121	rs28834970	<i>PTK2B</i>	AD 1	1	8	2
8	27467686	rs9331896	<i>CLU</i>	AD 1	1	11	1
8	131075859	rs2045091	<i>ASAP1</i>	PSP	2	49	2
10	11720308	rs7920721	<i>ECHDC3</i>	AD 2	2	8	1
11	47557871; 47380340	rs10838725; rs3740688 *	<i>CELF1/ SPII- PU.1</i>	AD 1&2	2	36	5
11	59923508	rs983392	<i>MS4A6A</i>	AD 1	1	177	4
11	85867875	rs10792832	<i>PICALM</i>	AD 1	1	64	0
11	121435587	rs11218343	<i>SORL1</i>	AD 1	1	3	0
12	21314281	rs7966334	<i>SLC01A2</i>	PSP	1	8	0
12	53788003	rs147124286	<i>SPI</i>	PSP	2	79	5
14	53400629; 53391680	rs17125944; rs17125924	<i>FERMT2</i>	AD 1&2	2	26	2
14	92926952; 92932828	rs10498633; rs12881735	<i>SLC24A4/RIN3</i>	AD 1&2	1	4	1
15	59045774	rs593742	<i>ADAM10</i>	AD 2	2	14	3
16	19808163	rs7185636	<i>IQCK</i>	AD 2	2	130	17
17	17q21.31	NA	<i>MAPT</i>	PSP	1	3482	194
17	61538148	rs138190086	<i>ACE</i>	AD 2	2	6	1
18	29088958	rs8093731	<i>DSG2</i>	AD 1	1	11	2
19	1063443	rs4147929	<i>ABCA7</i>	AD 1	1	8	0
19	19q13.32	NA	<i>APOE/TOMM40</i>	AD 1	1	640	37
19	51727962	rs3865444	<i>CD33</i>	AD 1	2	6	0
20	55018260; 54997568	rs7274581; rs6024870	<i>CASS4</i>	AD 1&2	2	17	4
21	28156856	rs2830500	<i>ADAMTS1</i>	AD 2	2	1	1

Source GWAS for tested loci and variants. PSP = Chen *et al.*, 2018 ; AD 1 = Lambert *et al.*, 2013; AD 2 = Kunkle *et al.*, 2019.

* Locus re-mapped in AD GWAS 2. Lead SNPs are not in LD. Both lead SNPs were tested here.

Table 1: Description of GWAS loci and variation tested in this study. Shown are the locus lead SNPs and annotated genes as described in the listed GWASs. For each locus: the MPRA stage in which it was tested, the number of variants tested per locus, and the number of variants with significant allelic skew (FDR $q < 0.01$; SigVars) ultimately identified.

Supplementary Materials for

Functional regulatory variants implicate distinct transcriptional networks in dementia

Yonatan A. Cooper ^{1,2,3}, Jessica E. Davis ⁴, Sriram Kosuri ⁴, Giovanni Coppola ^{3,5}, Daniel H.
Geschwind ^{1,6,7,8,*}

Correspondence to: dhg@mednet.ucla.edu

Supplementary Materials include:

Materials and Methods
Supplementary Text
Figs. S1 to S8
Tables S1 to S3
Captions for Data S1 to S3

Other Supplementary Materials for this manuscript include the following:

Data S1 to S3
S1: Supplemental Data
S2: Source Data
S3: Supplemental Information - Methods

Materials and Methods

Identification of candidate variants

MPRA Stage 1: We selected all genome wide significant ($p < 5 \times 10^{-8}$) variants from an AD and PSP GWAS (6)(10). We then identified all variants with a MAF $> 5\%$ in LD ($r^2 > 0.8$) with these variants in Europeans (CEU + FIN + GBR + IBS + TSI; 1000 Genomes Phase 3) using the LDProxy tool accessed through the LDlink API. We subsequently filtered out indels, multiallelic, and coding variants. Both alleles of each variant were centered in 162 bp of genomic context (hg19/37) using the biomaRt (2.44.0) and BSgenome (1.56.0) packages in R (4.0.0) to create oligos. We then removed oligos containing KpnI, MluI, SpeI, and XbaI restriction sites needed for library cloning, leaving 5,223 total variants. Finally, we appended 5' (CTGAGTACTGTATGGGCGACGCGT) and 3' (GGTACCGACAAAAGTGTCAACTGT) PCR adaptor sequences to each oligo and synthesized the library on an Agilent 15k 210-mer array.

MPRA Stage 2: We replicated a selection of 186 SigVars (FDR adjusted $q < 0.01$) identified in MPRA 1 and 140 negative control variants. For 212 variants we also created oligos: 1) in reverse complement orientation, 2) with the variant located in the bottom third of the genomic context, and 3) in the reverse orientation (Supplemental Fig. 3c; Supplementary Text). Furthermore, in MPRA 2 we attempted to re-assess variants that dropped out of MPRA 1 (defined below). Finally, we assessed the lead SNPs from additional significant loci from two AD GWAS (6, 7), as well as 4 PSP genome-wide suggestive loci (10) (Table 1). LD partners were identified as above, constituting an additional 483 variants. The final MPRA 2 library was synthesized in duplicate on an Agilent 7.5k 210-mer array. All tested loci and oligo sequences for both MPRA stages are listed in Supplemental Information – Methods.

Custom MPRA vector design

The pAAV-stop-MCS-bGH plasmid was created as follows: The multiple cloning site (MCS) and bGH-polyA sequence from the Donor_eGP2AP_RC plasmid (Addgene #133784) was cloned into the pAAV.CMV.PI.EGFP.WPRE (Addgene #105530) backbone using NheI and SphI restriction sites. To prevent transcriptional readthrough from the AAV ITR, this vector was re-cut at the NheI site and the transcriptional insulator element from pGL4.23 (Promega, E8411) was inserted using Gibson Assembly (78). The pMPRAdonor2-eGFP plasmid was created by cloning the eGFP open reading frame into the pMPRAdonor2 plasmid (Addgene #49353) digested with NcoI and XbaI restriction enzymes. Full plasmid sequences are provided in Supplemental Information - Methods.

Library Construction

Sequences for primers described below are listed in Supplemental Information – Methods.

Step 1: We amplified and attached 20 bp degenerate barcodes to the oligo library by emulsion PCR (Chimerx, 3600-01). We performed four 50 uL PCR reactions with individual mixtures containing: 2 pmol of library, 1 uM of barcode_new_F primer, 1 uM of barcode_N_R primer, 200 uM dNTPs, 0.25 mg/mL acetyl-BSA (Thermo Fisher Scientific, AM2614), and 2 U of Phusion Hot Start II DNA Polymerase (Thermo Fisher Scientific, F549S) in 1X HF buffer. Thermal cycle conditions were: initial denaturation for 1 min at 95°C, followed by 20 cycles of 10 sec at 95°C, 20 sec at 61°C, and 20 sec at 72°C (2.5°C/sec ramp rate), followed by a final extension for 5 min at 72°C. Emulsions were broken with butanol, pooled, and purified per manufacturer's instructions on spin columns. The amplified library and the pAAV-stop-MCS-

bGH plasmid were digested overnight using SpeI-HF and MluI-HF enzymes and purified using Streptavidin M-270 Dynabeads (Thermo Fisher Scientific, 65305) or gel purification respectively (28704, Qiagen). An 80 μ L T7 ligation reaction (NEB, M0318S) containing 200 ng of digested plasmid and 37.7 ng of library was performed followed by cleanup and electroporation into DH5 α electrocompetent cells (NEB, C2989K). Transformed bacteria were pooled, serially diluted, and plated overnight at 37°C. Colonies from the dilution plate containing the number of bacterial colonies approximating 50-fold library coverage were collected and grown, followed by Maxiprep library extraction (Thermo Fisher Scientific, K210016).

Barcode mapping: Sequencing was performed to create a lookup table mapping barcodes to oligos. Oligo-barcode sequences were amplified from 2 ng of plasmid using flanking PCR primers (BC_map_P5_Rev and BCmap_P7_For) that added P5 and P7 adaptor sequences. Amplicons were sequenced by the UCLA TCGB core using an Illumina NextSeq 550 system (PE 2x150 bp) using custom Read 1 (BCmap_R1Seq_Rev) and Read 2 (BCmap_R2Seq_For) sequencing primers. Reads were merged using the BBMerge tool (79) and barcodes filtered and assigned to oligos using a python script. Briefly, reads that did not perfectly match library oligos were discarded. Barcodes represented by fewer than three reads were dropped. Ambiguously mapped barcodes were then filtered as follows: We bootstrapped an empirical distribution of oligo Levenshtein distances (python-Levenshtein 0.12.0) to determine a cutoff score (1st percentile of distances). We then discarded barcodes where any pairwise read distance was greater than this cutoff score. The MPRA_barcode_mapping.py python script is provided: <https://github.com/ycooper27/Tauopathy-MPRA>.

Step 2: The MinP-eGFP fragment was amplified from the pMPRA_{donor2}-eGFP vector using Amp_minPLuc2_For and Amp_minPLuc2_Rev primers and both the plasmid library and fragment were sequentially digested using KpnI-HF and XbaI enzymes. These were used in a T7 ligation reaction containing 200 ng plasmid and 125 ng fragment. The ligation product was transformed into DH5 α competent cells followed by plasmid isolation. The final library was configured with the 162 bp oligo upstream of the minimal promoter and the 20 bp barcode located in the 3' UTR of the eGFP transcript (Fig. S1).

Massively Parallel Reporter Assay

MPRA was performed with 6 biological replicates each consisting of ~8 million HEK293T cells. We transfected 10 μ g of library plasmid per replicate using Lipofectamine 3000 (Thermo Fisher Scientific, L3000008). 24 hours post-transfection, cells were dissociated, pelleted, and washed with DPBS. Replicates were lysed in 1 mL of RLT buffer and homogenized using QIAshredder columns (Qiagen, 79654). Total RNA was extracted and eluted in 100 μ L RNase-free water using an RNeasy Mini Kit (Qiagen, 74104) with on-column DNase I digestion (Qiagen, 79254). We then extracted mRNA from 75 μ g total RNA per sample using a Dynabeads mRNA Purification Kit (Thermo Fisher Scientific, 61006). Residual DNA was removed from 1 μ g of mRNA using ezDNase (Thermo Fisher Scientific, 11766051) followed by reverse transcription using a custom primer (Lib_Hand_RT) and the SuperScript IV First Strand Synthesis Kit (Thermo Fisher Scientific, 18091050). RT was performed for 80 min at 52°C. We then amplified 10 μ L of cDNA in 100 μ L PCR reactions using NEBNext Ultra II Q5 Master Mix (NEB, M0544S) and Lib_Seq_eGFP_F2 and Lib_Hand primers for either 8 or 3 cycles (MPRA 1 and 2 respectively). Likewise, 5 replicates of 25 ng plasmid DNA (MPRA 1) or 4 replicates of

5 ng plasmid (MPRA 2) were amplified in 50 uL reactions using Lib_Seq_eGFP_F2 and Lib_Hand_RT primers. PCR products were purified using 0.6X-1.2X KAPA Pure Beads (Roche, KK8000) and then further amplified using P5_seq_eGFP_F2 and P7_Ind_#_Han primers for 7 PCR cycles to add P5, P7, and unique 8 bp Illumina index sequences. Following SPRI cleanup, amplicons were sequenced using an Illumina NextSeq 550 (1x20) or NovaSeq 6000 SP flow cell (SR 1x26 cycles) with 5% PhiX spike-in and custom Read 1 (Exp_eGFP_Seq_F2) and Index (Exp_Ind_Seq_P) primers.

MPRA data analysis

Preprocessing: Raw reads were trimmed to only contain the 20 bp barcode and aligned to the previously generated oligo-barcode lookup table to create a barcode count matrix. Barcode reads that did not perfectly match were discarded. The MPRA_BC_counter.py mapping script is provided: <https://github.com/ycooper27/Tauopathy-MPRA>.

Barcodes were filtered such that at least 1 count had to be detected in every RNA replicate and at least 5 counts detected in each DNA replicate (to ensure stability of the log-ratios). A pseudocount of 1 was added to each barcode count, which was then normalized to sequencing depth to create counts per million reads per replicate (CPM). We then computed $\log_2$ RNA/DNA ratios for each barcode (DNA = median count across plasmid replicates), which were then quantile normalized between replicates using the preprocessCore (1.50.0) package. Variants with fewer than 5 unique barcodes for either allele were removed from further analysis.

Oligo activity measurements: To calculate transcriptional activity scores for each 162 bp oligo, allele-level summary statistics were computed as the median of the $\log_2$ barcode ratios. This

value was then averaged between reference and alternate alleles to create an activity score for each oligo. To determine significance, this activity score was compared to the median activity value for the entire library using a one-sample Mann-Whitney-U test (two-tailed; $n = 6$ replicates), which was subsequently adjusted for multiple comparisons. “Active” elements for subsequent analyses had an increased RNA/DNA ratio compared with the library median at a Bonferonni adjusted $p < 0.05$ (or FDR adjusted $q < 0.01$ where described; Benjamini Hochberg method).

SigVar calculations: To identify variants with significant allelic skew (SigVars), $\log_2$ barcode ratios were combined across all 6 replicates by taking the median value. For each variant, a two-way Mann-Whitney-U test comparing barcode counts between each allele was used to identify allelic skew. SigVars were defined at an FDR threshold $q < 0.01$ (FDR adjustment, Benjamini-Hochberg method; further discussed below in Statistical Reporting). MPRA $\log_2$ effect sizes were defined as the median summed normalized barcode count for the alternate allele - reference allele.

Quality Control: Intra-experiment barcode reproducibility was defined as the mean pairwise correlation of each normalized barcode (RNA/DNA) count across all technical replicates. Allele correlation was determined by first finding the median normalized barcode count for each allele followed by determining mean pairwise correlation across all technical replicates (Both Pearson’s r and Spearman’s ρ computed). Between experiment correlations for reference allele activity scores and variant effect sizes were also determined for 326 variants replicated in MPRA 2 estimate inter-experiment reproducibility.

Variant Dropout: Variants were excluded from analysis if we were unable to obtain activity measurements from at least 5 unique barcodes for both alleles (Fig. S8a; Supplementary Text). Of the 5,706 unique SNPs tested in both MPRA stages, 366 did not meet this inclusion threshold in either stage. We then compared sequence features from these “dropout” vs the rest of the “included” oligos. We used the SeqComplex perl module to compute GC content, CpG skew, and sequence complexity metrics (<https://github.com/caballero/SeqComplex>) (80). Sequence entropy was calculated using the RNAfold tool from the ViennaRNA 2.0 software suite (81). We also generated a custom python script to identify the longest runs of mono- or di-nucleotide repeats within each oligo (script provided: <https://github.com/ycooper27/Tauopathy-MPRA>). Outputted scores were compared between “dropouts” and all “tested” oligos using a two-sided Mann-Whitney-U test. All sequence scores are provided in Source Data. For visualization (Fig. S8) we took a random sample of 366 “tested” variants.

Power Analysis: We performed a power analysis to determine the sensitivity of our assay at different barcode complexities (Fig. S8b). We first determined empirical SigVar effect sizes from our combined study, which were binned into percentiles. We also determined an empirical assay standard deviation by taking the average standard deviation of the normalized barcode counts for all alleles that passed filter in both studies. We then performed a power analysis using the power.t.test function from the stats package in R, using the empirical standard deviation, an alpha threshold of $p < 0.01$ and a “two.sided” hypothesis test. The analysis was performed using different percentiles (0, 20, 40, 60, 80th) of empirically computed effect sizes, as well as all integer n (i.e. # unique barcodes per allele) between 5-100, and plotted using ggplot2.

Bioinformatic Analyses

Jaccard Index Calculations: We download DNase I hotspot files from the ENCODE project server (<https://www.encodeproject.org/>) (82) for brain cell types and tissues (cell types, accessions listed in Supplemental Information – Methods). Pairwise Jaccard indices between all samples were calculated using the Jaccard tool from BEDTools (2.29.2) (83) and plotted as a heatmap (Fig. 1b). Biological replicates (rep #2) were then discarded to avoid artificial inflation (leaving only one sample per cell type), and the average pairwise Jaccard index for each cell type with all other cell types was computed. GWAS overlap (Fig. 1c) was calculated for a given cell type by first intersecting all tested GWAS variants with that cell type’s DNase I peaks. Then, only the DNase-overlapping variants were intersected with DNase I peaks from the other cell types (pairwise) and the mean proportion shared was computed.

Chromatin annotation enrichment: We downloaded narrowPeak files for HEK293 DNase-seq, histone ChIP-Seq, and TF-ChIP-seq marks (accessions listed in Supplemental Information – Methods) from the ENCODE project server (82). We then determined overlap between these marks and MPRA “active” and “repressive” elements using the GenomicRanges R package (1.40.0) assuming a minimum of 1 bp overlap between the 162 bp oligo and the chromatin mark. Enrichment was calculated for active or repressive elements against a background set of all other tested oligos using a Fisher’s exact test, with log₂ odds ratios and 95% confidence intervals reported. Active and repressive scores are provided in Source Data.

TFBS analysis: Predicted TFBS enrichment for active elements was calculated using the HOMER software suite (4.11) against a background set of all other oligos, after prefiltering for oligos containing reference alleles (Fig. 2e). In Fig. 4e, SigVar TFBS disruption was scored for all variants using the SNPS2TFBS webtool (43) and an enrichment odds ratio for TFBS-disrupting variants was calculated using Fisher's exact test. For SigVars predicted to disrupt TFBSs, we correlated allelic skew effect sizes from our MPRA with predicted TFBS disruption scores. Sites with a TFBS score of 0 (denoting poorly defined scores) were discarded and the score with the max absolute value was chosen for sites with multiple predicted disruptions. We performed a similar analysis for SNPs from another large MPRA dataset (Fig. 4f) (28). As this dataset characterized variants embedded in genomic context from both positive and negative strands, we used the max observed effect size. We also partitioned SigVars into those that were derived from AD or PSP GWAS (111 or 209 variants respectively; 17q21.31 variants were considered PSP), and re-ran them through the SNPS2TFBS algorithm (Source data). The output was filtered for TFBSs with at least two predicted disruptions and enrichment p-values were then FDR-adjusted (BH method). Only TFBSs significantly ($q < 0.05$ or $q < 0.1$) enriched for disruption by SigVars were plotted using ggplot2, with disruption counts, $\log_2$ enrichment odds ratios, and unadjusted $-\log_{10}$ p-values shown (Fig. 4g; Fig. S6).

We also tested whether HEK293T cell TF abundance correlates with MPRA allelic skew. To do so, we scored all variants tested in MPRA 1 using the SNPS2TFBS software. For each variant predicted to disrupt a TFBS, we found the corresponding TF's normalized gene expression in HEK293T cells using RNA-seq data from the human protein atlas (84) (accession in Supplemental Information – Methods). We then found the Spearman's correlation between TF expression and the MPRA absolute value $\log_2$ fold change.

For the 6 TFs significantly disrupted in PSP (at $q < 0.05$) or 14 TFs at $q < 0.1$, we created a protein-protein interaction network using the STRING (v11) webtool and standard parameters (Fig. 4g; Fig. S6) (77). We then computed empirical SP1-connectivity p-values for this network using resampling as follows: We determined protein-protein interactions for all TFs annotated by the SNPS2TFBS tool (165 TFs total; output in Source Data – Fig 4g.3) using STRING (<https://string-db.org>). We then found the number of edges between SP1 and either 5 or 13 additional randomly sampled TFs, repeating this procedure 10,000 times to create a distribution, which was compared to the true number of PSP-network edges (4 or 8 respectively) to generate a SP1-connectivity p-value. We then identified all sites among the 209 PSP-SigVars predicted to disrupt binding of these 6 TFs (union of SNPS2TFBS and motifbreakR annotations; Supplemental Table 1) and found all genes within +/- 10 kb of these sites, which we called “target genes” (Fig. 4h). We then annotated the network composed of these TFs and their targets using single cell RNA-seq data from human M1 cortex (© 2010 Allen Institute for Brain Science. Allen Human Brain Atlas. Available from: <https://portal.brain-map.org/atlas-and-data/rnaseq/human-m1-10x>) (45). We annotated each network gene as belonging to the corresponding cell type with the highest trimmed mean gene expression value, and created a boxplot displaying the number of annotated genes per cell type (Fig. 4i).

We also tested whether TFBS disruption was predictive of MPRA allelic skew. For this analysis, we considered variants derived from the external MPRA dataset (28). Only variants with computed allelic skew values were considered. We used the maximum $-\log_{10}$ p-value for variants tested in multiple configurations and also filtered out variants missing valid rsIDs. We then performed FDR adjustment (BH method) on the MPRA-determined allelic skew p-values, and labeled SigVars at $q < 0.01$ thresholds. All variants were scored for TFBS disruption using

the motifbreakR package (2.2.0) utilizing the HOCOMOCO v10 TF binding model (47, 85) (“strong” effect, and binding threshold of $p < 1 \times 10^{-4}$). Variants that could not be identified in dbSNP v144 were not scored and were subsequently discarded. We then computed the positive predictive value between predicted TFBS disruption and SigVar labels. We also downloaded RNA-seq data for K562 cells from the human protein atlas (84), and identified the top 200 TFs (by normalized counts) expressed in these cells. We then filtered for “strong” predicted TFBS disruption only for motifs corresponding to these top expressed TFs, and then re-computed PPD.

Finally, we tested whether disruption of particular TFBSs predicted MPRA allelic skew (Fig. S5). We again only considered K562 expressing TFs. For each collection of variants predicted to disrupt a particular TFBS, we found the proportion that were also annotated as SigVars (SigVars defined at $q < 0.01$). We then assessed the statistical significance of these proportions for each TF by performing a one-tailed binomial test against the background SigVar probability in the overall dataset ($\text{prob} = 0.115$), leaving a p-value which was FDR adjusted. Enrichments are shown in Source Data.

Comparison with computational predication algorithms: We scored all tested variants using the LINSIGHT, CADD, CATO, and GWAVA algorithms (39–42), using each algorithm’s precomputed scores. When an algorithm provided multiple scores per variant (particularly LINSIGHT), scores were averaged. Individual variant scores are provided in Source Data. First, we correlated (Spearman’s rho) MPRA effect-sizes with computational predicted scores. We then labeled MPRA SigVars at FDR-adjusted $q < 0.01$ thresholds and calculated AUC for each algorithm using the ModelMetrics (1.2.2.2) package. We then performed binary classification on all scored variants by positively labeling MPRA SigVars ($q < 0.01$) and an equivalent number of

the top scoring variants from each algorithm before calculating pairwise Cohen's Kappa using the ModelMetrics package. These analyses were replicated using another published MPRA dataset (28) as above.

SigVar functional annotations: SigVars from this study were annotated for TFBS disruption and overlapped with functional brain annotations (Supplemental Table 1). We calculated TFBS disruption using both the motifbreakR package using the HOCOMOCO v10 TF binding model (filtered for a binding threshold of $p < 1 \times 10^{-4}$ and “strong” predicted effects) as well as the SNPS2TFBS webtool (43). Additionally, published enhancer and promoter annotations for sorted microglia, neurons, astrocytes, and oligodendrocytes were downloaded and converted to bed files using the ucsc-bigbedtobed tool and overlapped with SigVars using BEDTools intersect (36, 83). We also identified SigVars overlapping high-confidence multi-tissue enhancers defined by the HACER database (59). Promoter/Enhancer accessions are provided in Supplemental Information – Methods.

LD clustering: To calculate LD between SigVars within the 17q21.31 locus, we downloaded chr17 VCF files from the 1000 Genomes FTP server. The VCF was subsetting for 90 unrelated individuals of CEU ancestry and reformatted to get cumulative allele frequencies using PLINK 1.9 (86). Genotype files were loaded into R and used to create LD clusters for SigVars within 17q21.31 using the clqd function (CLQDmode = “maximal”) from the gpart R package (1.6.0) (48).

Cell Culture

We obtained HEK293T (CRL-3216), THP-1 (TIB-202), and SH-SY5Y (CRL-2266) cell lines from ATCC. HEK293T cells were cultured in DMEM containing GlutaMAX (Thermo Fisher Scientific, 10566016) supplemented with 10% FBS and 1% Sodium Pyruvate (11360070). THP-1 cells were cultured in RPMI-1640 medium (ATCC 30-2001) supplemented with 10% heat-inactivated FCS and 10 mM HEPES buffer. THP-1 differentiation was performed through addition of 20 ng/mL of PMA (Millipore Sigma, P1585) to the culture media for 48 hours. SH-SY5Y cells were cultured in 1:1 EMEM/Ham's F12 (ATCC 30-2003 / Thermo Fisher Scientific, 31765035) supplemented with 15% HI-FCS, 1% Sodium Pyruvate, and 1% MEM-NEAA (Thermo Fisher Scientific, 11140050). For differentiation, SH-SY5Y cells were gently passaged and plated on Poly-D Lysine coated plates. After 24 hours culture media was replaced with a differentiation media composed of Neurobasal A (Thermo Fisher Scientific, 10888022), GlutaMAX, B27 Supplement (17504044), and 10 uM Retinoic Acid (Millipore Sigma, R2625). Cells were differentiated for 5 days with half-media replacement every 48 hours.

Induced Pluripotent Stem Cells (iPSCs) used in this study were previously documented (87) and kindly provided by Dr. Li Gan in accordance with the UCLA TDG guidelines. iPSCs were differentiated into mature astrocytes for 120 days as previously described (88), and were maintained post-differentiation in DMEM supplemented with 1% Sodium Pyruvate, 10% HI-FCS, and 1x N2 supplement (Thermo Fisher Scientific, 17502048) until use.

CRISPR experiments

We excised enhancers containing rs636317, rs13025717, rs6064392, rs9271171, and rs7920721 as follows: Pairs of guide RNAs targeting upstream (5') and downstream (3') flanking sequences were designed and cloned into LentiCRISPRv2-GFP (Addgene #82416) and

LentiCRISPRv2-mCherry (#99154) respectively using the BbsI restriction site (gRNA sequences in Supplemental Information – Methods). Lentiviral particles were produced in HEK293FT cells by triple transfection as previously described (<https://www.addgene.org/protocols/lentivirus-production>), concentrated using Lenti-X (Takara Bio, 631232), and resuspended in DPBS. Guide pairs were then screened for cutting efficiency resulting in two pairs targeting rs13025717 (gRNA #s: 11 + 12, 11 + 14), one pair targeting rs636317 (15 + 16), one pair targeting rs6064392 (25 + 26), three pairs targeting rs9271171 (3 + 4, 5 + 4, 5 + 6), and one pair targeting rs7920721 (21 + 20). Guide pairs or scramble gRNA control lentiviruses (MOI ~ 0.5) were then used to infect 80% confluent 6-well plates of SH-SY5Y cells or t25 flasks of THP-1 cells. Culture media was replaced 16 hours later and cells were expanded for 5 days post-infection. Cells were sorted at the UCLA BSCRC flow cytometry core to isolate ~500,000 GFP+/mCherry+ cells per replicate, and were subsequently differentiated. Ultimately we excised a 240 or 374 bp region containing rs13025717 in SH-SY5Y cells, a 430 bp region containing rs636317 in THP-1 cells, a 382 bp region containing rs6064392 in THP-1 cells, a 1065 or 682 bp region containing rs9271171 in THP-1 cells (as well as a 1005 bp region in iPSC-derived astrocytes), and a 473 bp region surrounding rs7920721 in THP-1 cells.

For each replicate we collected total RNA and genomic DNA using the AllPrep DNA/RNA Mini Kit (Qiagen, 80204). CRISPR-mediated removal of the target enhancer was assessed by amplifying the target region of gDNA by PCR (genomic PCR primers: Supplemental Information - Methods) and verifying strong representation of the truncated allele via gel electrophoresis. cDNA was reverse transcribed using SuperScript IV, Oligo(dT)₂₀ primer, and 300ng of total RNA. We performed qPCR using the KAPA SYBR FAST Kit (Roche, KK4600), 500nM qPCR primers (qPCR primers: Supplemental Information - Methods) and a Roche

LightCycler 480. Relative transcript abundance was quantified using the $2^{-\Delta\Delta CT}$ method (89) normalized to the geometric mean of ACTB and GAPDH reference genes.

Statistical Reporting

Statistical analysis was performed using the stats package in R. All hypothesis testing was two-sided except for the binomial test used in Supplemental Fig. S5. Unless otherwise stated, all enrichment analysis was performed using a Fisher's exact test. MPRA allelic skew multiple testing correction was performed as follows: Variants from MPRA 1 were combined with additional unique variants tested in MPRA 2 (total 5340 variants) and Mann-Whitney-U p-values were FDR-adjusted (BH method). SigVars were called at a threshold of $q < 0.01$. To assess variant reproducibility, variants replicated, *i.e.* re-tested (320 total) in MPRA 2 were considered separately, and assigned significance at a Bonferroni-adjusted $q < 0.05$.

Data Visualization

Variant genomic annotations were determined and plotted using the annotatr (1.14.0) Bioconductor package and ggplot2 from the tidyverse collection (1.3.0). Heatmaps were generated using the pheatmap R package (1.0.12). The circle Manhattan plot was created using the CMplot R package (<https://github.com/YinLiLin/R-CMplot>). ROC plots were visualized using the plotROC (2.2.1) extension and ggplot2 (90). 17q21.31 LD plots were created using the BigLD function from the gpart package (48). Genomic tracks and chromatin annotations were visualized using the pyGenomeTracks python module (91). All other data were visualized using ggplot2 with the GGalley package extension.

Supplementary Text

Massively Parallel Reporter Assay Technical Considerations

In this supplemental note, we provide further discussion and analysis of factors observed to impact MPRA experimental outcomes and technical performance. Consideration of these parameters may aid in the design and implementation of future studies involving complex library construction and massively multiplexed assays.

MPRA reproducibility

Because of our two-staged experimental design, we took the opportunity to characterize the true reproducibility of our assay. We identified 326 variants in the first MPRA stage, including 186 with significant transcriptional skew between alleles (SigVars; FDR adjusted $q < 0.01$) to be re-tested in the second stage. Importantly, although library design stayed identical, oligo barcoding, cell culture, sequencing, and other technical factors were distinct between stages, thereby providing an informative estimate of the impact of technical confounding variables. Fortunately, we found that our assay was highly reproducible, with measurements of allele transcriptional efficacy, and relatedly, MPRA-determined effect sizes, highly correlated between experiments ($r = 0.98$ and 0.94 respectively, both $p < 2 \times 10^{-16}$; Fig. 3a). This suggests that MPRA-determined measurements of transcriptional efficacy are highly precise and that variant ordering (e.g. prioritization) on the basis of effect size is robust. Significantly, reproducibility is not an inherent feature of SigVars or highly expressed variants in particular. Upon inspection of the 140 re-tested variants without significant allelic skew ($q > 0.01$), we again confirmed high correlation between stages ($r = 0.96$, $p < 2 \times 10^{-16}$)

Interestingly, 152 of 186 (82%) re-tested MPRA 1 SigVars remained significant in stage 2 (replication threshold; Bonferroni $q < 0.05$; Fig. 3a), which is somewhat lower than might be

expected considering the original conservative FDR threshold of $q < 0.01$ used to determine significance in stage 1 and the near perfect correlation of effect sizes observed between experiments. In comparing the 34 non-replicating and 152 replicating variants, the non-replicates had a *higher* MPRA stage 1 median barcode complexity (least complex allele; 94 vs 67 barcodes), and a *lower* MPRA stage 2 barcode complexity (45 vs 65). Thus, variants that failed to replicate had a dramatic reduction in barcode complexity in the replication stage (delta barcodes = -49 vs -3 barcodes). Concurrently, non-replicating variants had a significantly lower mean absolute effect size (MPRA stage 1 Log₂ FC: 0.36 vs 0.72, $p = 2.6 \times 10^{-9}$; two-sided Mann-Whitney-U test). We next determined power to detect significant allelic skew for a range of percentiles of empirical MPRA effect sizes at different levels of barcode complexity (Fig. S8b; Methods). Interestingly, the threshold at which 80% power is achieved for the 40th percentile of empirical effect sizes is at a barcode complexity of 42 barcodes per allele, with a steep drop-off in power at decreasing effect sizes and barcode complexities. Thus power to detect SigVars of small effects are highly contingent on high (>40 barcodes/allele) barcode complexity, and suggests that while MPRA is highly precise in determining allele effect sizes (even at low complexities ~5 barcodes, Data not shown), identification of allelic skew *significance* will be somewhat impacted by stochastic fluctuations in barcode complexity due to PCR.

Effects of oligo configuration on MPRA performance

We also tested the effects of library construction parameters on SigVar detection and reproducibility, using the SigVars identified in MPRA stage 1 as a “gold standard” test set. Specifically, we tested the impact of placing the oligo in the reverse complement (RC) orientation in the MPRA vector as well as the effect of placing the variant in the bottom third (as

opposed to the middle) of the 162 bp genomic context. We used the oligo in reverse orientation as a negative control as this maintained identical nucleotide composition (Fig. S3c).

Interestingly, SigVars had strong but imperfect correlation with their RC ($r = 0.69$, $p < 2 \times 10^{-16}$) and lower third ($r = 0.78$, $p < 2 \times 10^{-16}$) counterparts, suggesting modest effects of oligo orientation and distal sequence context on MPRA activity. As MPRA activity measures are highly precise and reproducible (discussed above), these effects of oligo configuration are likely biological and not due underlying assay noise. Placing oligos in the reverse orientation completely abolished activity as expected (all $p > 0.5$; Fig. S3c). These findings can inform design considerations for future massively parallel screens.

Technical factors influencing variant drop-out

We initially assumed that the observed 9% variant drop-out during stage 1 was random bottlenecking due to library construction and therefore attempted to re-test missing variants in stage 2. However 346/491 of these variants also failed to pass QC in stage 2, suggesting that a large proportion of drop-outs were due to systematic amplification failure during library construction or strong transcriptional repression. Indeed, of the vast majority of variants that dropped out specifically during MPRA 1 library construction dropped out again during MPRA 2 library construction, suggesting PCR amplification failure. We assessed sequence features of the 366 unique variants that failed to pass quality control (QC) for both MPRA 1 and 2 and found that missing variants were more likely to contain GC content greater than 75% or less than 25% (GC extremes) and had increased rates of CpG skew. Additionally, drop-out variants had significantly lower mean sequence complexity (measured by Shannon's entropy, compressibility,

linguistic complexity measures, *etc.*; Methods), and were more likely to contain long regions of repetitive sequences such as CpGs, dinucleotide repeats, or polyA tracts (Fig. S8a).

Library expression impacts assay performance

It has been previously reported that MPRA performance, as defined by the precision and reproducibility of measurements of library transcriptional efficacy, is dependent on high levels of library expression (33). This is particularly true for assays measuring variant function, which attempt to identify subtle changes in transcriptional efficacy between alleles (i.e. detect small effect size events). Three critical factors influencing expression of reporter libraries are: 1) the ability of library elements to drive transcription, 2) global transcriptional propensity of the particular cell-type, 3) transfection or infection efficiency into cells of interest (e.g. copy number per cell, cell transduction percentage). Unfortunately, assays testing variant allelic skew typically employ weak minimal promoters so as not to overwhelm the transcriptional impacts of single nucleotide substitutions, which are typically small. As a result, these assays have been exclusively performed in cancerous cell lines, which can be transfected at high efficiencies. Moreover, cancer cells, particularly those with c-Myc amplifications have been noted to have higher levels of overall transcription (known as transcriptional amplification) (92).

We confirmed the importance of library expression on assay performance by comparing four separate MPRA experiments: the two assays presented in this study as well as two unpublished assays. We took the average between-replicate correlation (Spearman's rho) of allele expression as the overall performance metric for each experiment. Significantly, each experiment required a differing number of PCR cycles to amplify an adequate amount of library cDNA for downstream sequencing (10, 15, 17, 18 total PCR cycles). We found an inverse

relationship between PCR cycle number and performance (mean rho; Fig. S8c), with a large performance drop off above 15 cycles. As PCR cycle number likely reflects underlying mRNA quantity and library expression, this confirms the importance of expression on performance. This observation represents a major technical obstacle towards implementing MPRA within difficult to transduce primary cell-lines, which may be desirable *in vitro* model systems that more closely recapitulate relevant biology than cancer lines.

Supplemental Fig. 1

Figure S1: Technical schematic for massively parallel reporter assay (MPRA). **(A)** We selected all bi-allelic noncoding genome-wide significant ($p > 5 \times 10^{-8}$) or locus lead SNPs and LD partners ($r^2 > 0.8$) from three GWAS (Methods). **(B)** The reference and alternate alleles of each variant were centered within 162 bp of genomic context (hg19) and synthesized on an Agilent array. **(C)** Oligos were amplified by PCR using primers complementary to flanking shared adaptor sequences. This step also attached 20 nucleotide random barcodes and restriction enzyme sites. **(D)** Oligos were digested and cloned into an expression library, followed by paired-end 2x150 bp sequencing to map barcodes to associated oligos. **(E)** A minimal promoter and eGFP gene was inserted between the oligo and barcode, with the oligo upstream of the minimal promoter and the barcode in the 3'UTR of eGFP. **(F)** The expression library was transfected into HEK293T cells for 24 hours. **(G)** following mRNA extraction, mRNA and plasmid barcodes were amplified, pooled, and sequenced using indexed single-end sequencing. **(H)** mRNA barcode counts were normalized to DNA counts, and the median normalized barcode count was taken as an activity summary score for each allele.

Supplemental Fig. 2

Fig. S2: MPRA 1 extended quality control metrics. **(A)** Density plot shows the distribution of uniquely mapped 20 nucleotide barcodes per individual allele (i.e. unique barcodes/allele; 9,464 total alleles) for the full MPRA 1 library, blue line = median (40 barcodes/allele). **(B)** Density plot of log₂ normalized barcode (BC) ratios. **(C)** Representative plot showing the correlation of log₂ normalized barcode counts between technical replicates (Pearson's $r = 0.57$; Spearman's $\rho = 0.50$; both $p < 2 \times 10^{-16}$). **(D)** Enrichment log₂ odds ratios (Fisher's exact test) of active and repressed elements (activity defined at FDR $q < 0.01$, BH method) within HEK293T ChIP-seq

peaks for both histone and TF marks. Error bars = 95% CI, *** FDR-adjusted $p < 0.001$ (BH method), ** $p < 0.01$, $p < 0.05$.

Supplemental Fig. 3

Figure S3: MPRA 2 quality metrics. **(A)** Barcode distribution for the 3,072 unique alleles tested in the full MPRA 2 library (blue line = median) with ii) representative plot showing correlation of normalized barcode (BC) counts between technical replicates (Pearson's $r = 0.83$, Spearman's $\rho = .79$, $p < 2 \times 10^{-16}$). **(B)** Panels show pairwise correlation of median $\log_2$ normalized barcode counts for all alleles passing filter ($n = 3,072$) between 6 technical replicates. Red line = OLS regression line of best fit, all $p < 2 \times 10^{-16}$. **(C)** i) Shows the experimental design for testing the effect of enhancer orientation and variant placement on MPRA activity. 212 variants identified from MPRA 1 were replicated in MPRA 2: Oligos were kept in the same orientation as the original (FDR), oligos were placed in the reverse complement orientation (Rev. Comp.), variant was placed in the lower third of the oligo rather than the middle (Lower Third), or the enhancer sequence was reversed (Reverse) as a negative control. Plots showing correlations between oligo orientations shown in ii), red line = OLS regression line of best fit, Pearson's correlation.

Supplemental Fig. 4

Figure S4: (A-B) All MPRA-tested variants were scored using four variant-effect prediction algorithms; LINSIGHT, CADD, GWAVA, and CATO. Violin plots show algorithm prediction scores for the top vs. bottom 5th percentile of ranked variants (rank determined by MPRA allelic

skew p-values) in the current study (**A**) and replication dataset (**B**). Mann-Whitney-U test. (**C-D**)

When comparing all tested variants, prediction algorithm scores correlated poorly with MPRA-determined allelic skew effect sizes in both our dataset (**C**) and the replication dataset (**D**). Black dashed line = OLS regression line of best fit. Pearson's r , all $p > 0.1$ (N.S.), except for CATO algorithm.

Supplemental Fig. 5

b

TF	Family	Motif	p-values	q-values
GABPA	Ets-related		3.5e-12	1.7e-10
RELA	NF-kB-related		1.4e-5	0.0003
ELF2	Ets-related		6.5e-5	0.0001
JUND	JUN		0.0005	0.006
ELK1	Ets-related		0.002	0.02
SP2	3-finger Kruppel-related SP - subfamily		0.002	0.02
SP1	3-finger Kruppel-related SP - subfamily		0.05	0.3

Figure S5: Transcription Factor Binding Site disruption is predictive of MPRA allelic skew for specific classes of TFs, including the ETS-related TF family. (A) Variants with significant allelic skew (SigVars) were defined at an FDR-adjusted threshold of $q < 0.01$ (Benjamini-Hochberg

method) for variants tested in a previously published large MPRA dataset. These variants were also assessed for predicted TFBS disruption (motifbreakR; Methods) filtering for motifs corresponding to the top 200 most highly expressed TFs in K562 cells. The histogram counts the number of TFBS-disrupting variants annotated for each TF, colored by whether variants are also MPRA SigVars (blue) or not (red). TFs for which disrupted binding significantly predicts MPRA allelic skew are highlighted in red, and also described further in (B). Determination of a significantly increased proportion of SigVars amongst variants predicted to disrupt specific TFBSs was assessed using a one-sided binomial test, with p-values and FDR-adjusted q-values shown (Methods).

Supplemental Fig. 6

Figure S6: Shows the $\log_2$ enrichments for significantly disrupted TFBSs (FDR $q < 0.1$, BH method), with SigVars from AD and PSP analyzed separately (color = $-\log_{10}$ enrichment p-values, size = # disrupted TFBSs; SNPS2TFBS). Inset: i) protein-protein interaction network from the STRING database for significantly disrupted PSP-TFs. Line thickness = strength of evidence. ii) Empirical distributions for expected PPI networkSP1 - connectivity generated by resampling (Methods). Red lines = observed PSP-TF network edges.

Supplemental Fig. 7

Figure S7: (A) CRISPR-mediated deletion of a small genomic region containing rs6064392 in differentiated THP-1 cells does not alter expression of *RTF2* (n=4/group; t(6) = 0.1; p = 0.94; two-tailed Student's t-test). (B-C) CRISPR-mediated deletion of genomic regions containing rs9271171 significantly increases expression of *HLA-DQA1* (n=4 group, n=8/condition, t(14) = -6.9; p = 7.5×10^{-6} ; two-tailed Student's t-test) in differentiated THP-1 cells (B) and decreases

expression of *C4A* in iPSC-derived astrocytes (CRISPR n = 3, control n = 2; $t(3) = 3.5$; $p = 0.04$; two-tailed Student's t-test). **(D)** CRISPR-mediated deletion of a small genomic region containing rs7920721 in differentiated THP-1 cells does not alter expression of *USP6NL* or *ECHDC3* (n=4/group; $t(6) = 0.1$ and 0.6 ; $p = 0.92$ and 0.6 respectively; two-tailed Student's t-test). Error bars = S.E.M.

Supplemental Fig. 8

Figure S8: Characterization of technical features impacting MPRA performance (Supplemental Text). (A) Assessment of sequence-level features of the 366 variants (“Dropout”) that failed to

pass quality control thresholds through both MPRA experimental stages (Methods). Violin plots display GC-content, CpG content, sequence complexity measures ($-\log_2$ compressibility and $\log_2$ Shannon's entropy), predicted secondary structure (calculated using RNAfold), and longest mono- or di-nucleotide repeat per tested oligo for the 366 dropout variants vs. a random sample of 366 variants that passed QC ("Tested"). Dropout variants were enriched for GC content extremes ($>75\%$ or $<25\%$), increased CpG skew, had decreased mean sequence complexity, and on average contained longer runs of nucleotide repeats (mean 13.2 vs 7.2 nucleotides) (2-tailed Mann-Whitney-U test). Predicted secondary structure did not differ for dropout variants ($p > 0.05$). **(B)** Simulated power analysis, using MPRA-determined empirical variance and different quantiles of empirical effect sizes, revealed that adequate power (0.8) to detect significant allelic skew at a broad range of effect-sizes is achieved at a barcode complexity of ~ 40 barcodes/allele. **(C)** Plots mean correlations (Spearman's rho) of inter-replicate allele expression for four separate MPRA experiments. Inter-replicate reproducibility is inversely proportional to the number of PCR cycles needed to amplify library mRNA before sequencing.

Table S1. SigVars with Annotations

rsID	chr_pos_A0_A1	Disorder	locus	Log2.FC	q.value	TFBS_1	TFBS_2	Enhancer	Promoter	cis_genes
rs2225332	chr1_180998570_A_G	PSP	2	2.12508	1.27E-17	NA	NA	NA	NA	STX6, MR1
rs6684455	chr1_181016529_A_G	PSP	2	0.82987	1.01E-12	NA	ELF3	NA	NA	MR1
rs6701713	chr1_207786289_A_G	AD	3	1.33975	4.27E-30	CTCF, ATOH1	CTCF, TAL1	NA	NA	CR1
rs2093761	chr1_207786542_A_G	AD	3	0.58514	3.15E-06	NA	NA	NA	NA	CR1
rs72740077	chr1_222114169_T_C	PSP	4	-0.23418	6.11E-04	NR2C2, PPARG_RXRA	PAX5	NA	NA	LINC02257
rs12123928	chr1_222121875_C_T	PSP	4	-0.20582	3.08E-03	NFE2L1_MAF G, PAX2	NA	NA	NA	LINC02257
rs12125368	chr1_222198672_C_A	PSP	4	-0.32564	2.13E-07	NA	NA	NA	NA	NA
rs7920721	chr10_11720308_A_G	AD	5	0.64992	1.76E-10	NA	NA	PU1	NA	NA
rs10838726	chr11_47568344_C_G	AD	6	0.16943	3.46E-03	NA	NA	NA	NA	CELF1
rs12287076	chr11_47606865_G_C	AD	6	-0.15062	5.68E-04	NA	NA	NEUN	NA	KBTD4, NDUFS3, FAM180B, C1QTNF4
rs11039332	chr11_47695840_G_A	AD	6	-0.70994	1.60E-06	NA	NA	NA	NA	AGBL2
rs12223593	chr11_47789082_T_G	AD	6	-0.46626	1.65E-08	NA	CEBPZ, NFYA, NFYB, NFYC, PBX3, YBX1	NA	LHX2, NEUN, OLIG2, PU1	FNBP4
rs7934481	chr11_47883337_C_T	AD	6	0.16539	7.32E-03	NA	NA	NA	NA	NA
rs11230153	chr11_59896901_G_T	AD	7	0.36167	8.99E-03	NR5A2	ESRRG, NFIA+NFIB+NFIC +NFIX, NR5A2, NR6A1	NA	NA	NA
rs493692	chr11_59909610_A_G	AD	7	-1.35534	1.93E-18	NA	FOXP2	NA	NA	NA
rs667897	chr11_59936979_A_G	AD	7	-1.01490	2.59E-13	NFE2_MAF	MAFK, NFE2, NFE2L2	NA	NA	MS4A6A
rs636317	chr11_60019150_C_T	AD	7	-1.95008	8.74E-21	CTCF, ATOH1	CTCF, MLXIPL, NEUROD1, TAL1, TCF12, TCF3	NA	NA	MS4A4E
rs11170528	chr12_53786242_G_A	PSP	11	-0.62823	1.96E-08	NA	E2F1, E2F4	NA	NA	SP1
rs10876450	chr12_53811034_T_C	PSP	11	-0.39697	9.70E-08	NA	ESR1, GATA5, MECOM, NR4A3	NA	NA	SP1, AMHR2
rs10876451	chr12_53813424_A_C	PSP	11	-0.23459	5.17E-05	NA	TCF4, TFE3	NA	NA	SP1, AMHR2
rs11170555	chr12_53824143_A_G	PSP	11	-0.40081	2.39E-07	NA	KLF3	NA	NA	AMHR2
rs4249196	chr12_53831848_T_C	PSP	11	-0.16027	1.45E-03	SOX10	AR	NA	NA	AMHR2, PRR13
rs7144029	chr14_53333616_C_A	AD	12	0.27128	1.01E-08	PRRX2	DLX3, HNF4G, LHX3, NR1I3, NR2F1, PDX1	NA	NA	FERMT2
rs17125924	chr14_53391680_A_G	AD	12	0.61092	9.07E-10	NA	ELF3, ETV7	NA	NA	FERMT2
rs36026988	chr14_92938382_T_C	AD	13	0.40229	4.77E-08	NA	NA	NA	NA	SLC24A4
rs6494035	chr15_59015888_C_G	AD	14	0.41932	2.09E-03	NA	NA	NA	NA	ADAM10
rs572107	chr15_59055810_T_C	AD	14	0.24843	1.11E-03	NA	RORA, RORC	NA	NA	LOC101928725, MINDY2
rs602602	chr15_59057023_T_A	AD	14	-0.19028	2.52E-04	NR5A2, ZNF354C	NA	NA	NA	LOC101928725, MINDY2
rs9935063	chr16_19710702_A_G	AD	15	0.17838	8.48E-03	NA	ARNT2	NEUN, OLIG2	NA	VPS35L, KNOP1
rs11646891	chr16_19711351_C_T	AD	15	-0.24097	2.33E-03	NA	NA	LHX2, NEUN, OLIG2	NA	VPS35L, KNOP1
rs3829539	chr16_19722366_T_C	AD	15	-0.28982	8.02E-11	NA	NA	NA	NA	VPS35L, KNOP1, IQCK

rs4782272	chr16_19729016_G_C	AD	15	-0.43912	5.94E-08	NA	STAT3	NA	LHX2, NEUN, OLIG2, PU1	KNOP1, IQCK
rs4780824	chr16_19737425_A_G	AD	15	0.24396	1.22E-04	NA	SP1	NA	NA	KNOP1, IQCK
rs78731084	chr16_19742049_G_A	AD	15	-0.10990	4.76E-03	NA	SMAD2, SMAD3, SMAD4	NA	NA	IQCK
rs13331873	chr16_19749709_A_G	AD	15	0.24730	6.93E-04	ZNF263	GLIS3, PLAG1	NEUN	NA	IQCK
rs56265490	chr16_19787949_G_A	AD	15	-0.32192	4.26E-07	NA	NA	LHX2, OLIG2	NA	IQCK
rs9928277	chr16_19791663_C_T	AD	15	0.47338	1.96E-08	NA	NA	LHX2	NA	IQCK
rs12599372	chr16_19794321_G_T	AD	15	0.27426	2.86E-06	EGR2	FOXJ3	NA	NA	IQCK
rs11765075 7	chr16_19794325_G_T	AD	15	0.44717	3.54E-12	EGR2	FOXJ3, FOXM1	NA	NA	IQCK
rs7185335	chr16_19794374_A_G	AD	15	-0.36557	2.25E-06	NA	NA	NA	NA	IQCK
rs9932521	chr16_19797878_A_G	AD	15	0.18519	6.05E-03	NA	NA	LHX2, OLIG2	NA	IQCK
rs13332674	chr16_19803199_G_T	AD	15	-0.25561	3.41E-03	NA	PLAG1	NA	NA	IQCK
rs4782284	chr16_19809695_A_G	AD	15	0.64930	6.70E-04	NA	NA	NA	NA	IQCK
rs11865116	chr16_19814744_T_C	AD	15	0.36806	4.34E-07	SMAD2_SMA D3_SMAD4	TBX2, TBX3, TBX5, TCF12, TCF4	NA	NA	IQCK
rs11641590	chr16_19820241_G_A	AD	15	-0.18769	1.92E-05	NA	CEBPD, CEBPG	LHX2	NA	IQCK
rs13791262 1	chr17_42903795_G_A	PSP	16	0.74640	4.49E-19	ETS1, FLI1	ELF1, ELF3, ETV4, TFAP2D	NA	NA	GJC1
rs75042717	chr17_42923807_A_C	PSP	16	0.81537	9.81E-13	NFIC	GCM1, SMAD2	NA	NA	HIGD1B, EFTUD2
rs62065787	chr17_43105631_G_C	PSP	16	0.21793	5.47E-03	ZNF143	NA	NA	NA	DCAKD
rs6503419	chr17_43149988_T_C	PSP	16	0.31037	4.39E-07	NA	NA	NEUN	NA	NMT1
rs11656414	chr17_43337970_A_G	PSP	16	-0.35858	1.47E-03	NA	NA	NA	NA	MAP3K14-AS1, SPATA32, MAP3K14
rs9972936	chr17_43353677_G_A	PSP	16	-1.02752	2.21E-19	NA	NA	NA	NA	MAP3K14-AS1, MAP3K14
rs4792846	chr17_43367152_G_A	PSP	16	0.45423	2.11E-03	ETS1	IKZF1	LHX2, NEUN	NA	MAP3K14
rs11742161 2	chr17_43423162_G_A	PSP	16	0.46451	8.25E-04	NA	GFI1	NA	NA	NA, LRRC37A4P
rs12952708	chr17_43456728_C_T	PSP	16	0.31171	4.34E-06	TP63, ARNT_AHR	HEY2, TP53	NA	NA	NA, LRRC37A4P
rs12940792	chr17_43492357_C_T	PSP	16	0.42696	3.18E-03	NA	AIRE, AR, VDR	PU1	NA	ARHGAP27, LRRC37A4P
rs35884427	chr17_43495216_G_A	PSP	16	-0.57334	8.41E-06	NA	NR6A1	NA	NA	ARHGAP27, LRRC37A4P
rs1230082	chr17_43500344_G_C	PSP	16	0.79413	1.18E-11	NA	HAND1	NA	NA	ARHGAP27, LRRC37A4P
rs1230086	chr17_43503283_C_T	PSP	16	-0.20311	1.41E-03	ZNF263	NA	PU1	NA	ARHGAP27, PLEKHM1, LRRC37A4P
rs76344126	chr17_43503284_G_A	PSP	16	0.21101	1.06E-04	ZNF263	NA	PU1	NA	ARHGAP27, PLEKHM1, LRRC37A4P
rs7209501	chr17_43503294_A_C	PSP	16	0.29681	7.43E-05	ZNF263	MAZ	PU1	NA	ARHGAP27, PLEKHM1, LRRC37A4P
rs11702970 4	chr17_43519760_G_T	PSP	16	-0.33737	1.11E-09	NA	NA	NA	NA	ARHGAP27, PLEKHM1, LRRC37A4P
rs71369809	chr17_43539035_G_A	PSP	16	0.28217	1.76E-05	NA	NA	NA	NA	NA

rs62065403	chr17_43551523_T_C	PSP	16	0.44021	7.27E-07	NA	E2F1, E2F4	NA	NA	PLEKHM1, MIR4315-1, MIR4315-2, LRRC37A4P
rs4792817	chr17_43556734_A_T	PSP	16	0.37657	2.81E-09	NA	MEF2D, ZNF384	NA	NA	PLEKHM1, MIR4315-1, MIR4315-2, LRRC37A4P
rs55925547	chr17_43556807_T_C	PSP	16	-0.20070	1.34E-03	NA	CTCF, NFIA+NFIB+NFIC +NFI	NA	NA	PLEKHM1, MIR4315-1, MIR4315-2, LRRC37A4P
rs4792860	chr17_43556878_C_T	PSP	16	-1.88368	6.57E-04	NA	FOXC2	NA	NA	PLEKHM1, MIR4315-1, MIR4315-2, LRRC37A4P
rs11227579 3	chr17_43563304_A_C	PSP	16	0.59876	2.94E-06	HOXC9	NA	NA	NA	PLEKHM1, LRRC37A4P
rs11139225 1	chr17_43568928_T_A	PSP	16	-0.39496	6.04E-04	NA	IRF1, IRF2, IRF3	NA	LHX2, NEUN, OLIG2, PU1	PLEKHM1, LRRC37A4P
rs62066460	chr17_43574935_C_T	PSP	16	0.44404	1.99E-08	BHLHE40, ARNT	ARNTL, BHLHE40, BHLHE41, MAX, MYC, TFE3, TFEB, USF1	NA	NA	PLEKHM1, LRRC37A4P
rs62064393	chr17_43665343_T_G	PSP	16	-0.25897	2.89E-04	NA	NA	NA	NA	NA
rs1724397	chr17_43665809_C_T	PSP	16	1.35515	5.41E-21	NA	NA	NA	NA	NA
NA	chr17_43667922_G_C	PSP	16	0.50506	1.32E-08	NA	NA	NA	NA	NA
rs62064421	chr17_43667934_C_G	PSP	16	-1.01953	4.10E-28	NA	NA	NA	NA	MAPK8IP1P2
rs62064422	chr17_43668017_G_A	PSP	16	-0.51506	6.12E-07	NA	NA	NA	NA	MAPK8IP1P2
rs55836887	chr17_43668815_A_G	PSP	16	1.09787	4.90E-15	NA	NA	NA	NA	MAPK8IP1P2
rs62064431	chr17_43669704_C_A	PSP	16	-0.20592	7.92E-03	NA	NA	NA	NA	MAPK8IP1P2
rs1706731	chr17_43670696_G_C	PSP	16	0.53164	2.23E-04	NA	NA	NA	NA	MAPK8IP1P2
rs4471725	chr17_43683611_G_T	PSP	16	0.24944	3.37E-04	ARID3A	NA	NA	NA	MAPK8IP1P2
rs4485406	chr17_43683627_C_G	PSP	16	-0.21326	2.56E-03	NA	NA	NA	NA	MAPK8IP1P2
rs2906006	chr17_43685021_C_T	PSP	16	-0.72638	8.75E-12	NA	NA	NA	NA	MAPK8IP1P2
rs2942169	chr17_43711539_G_C	PSP	16	-0.24998	3.70E-03	NA	NA	NA	NA	LINC02210- CRHR1
rs422112	chr17_43728137_G_A	PSP	16	-0.17106	6.04E-04	NA	NA	NEUN	NA	LINC02210- CRHR1
rs647483	chr17_43729432_C_T	PSP	16	-0.18630	9.10E-03	NA	NA	NA	NA	LINC02210- CRHR1
rs1635298	chr17_43744344_T_A	PSP	16	0.21115	8.99E-03	PDX1, PRRX2	NA	NEUN	NA	LINC02210- CRHR1
rs1631850	chr17_43749942_A_G	PSP	16	0.93416	5.32E-35	EGR1, ARNT	NA	NA	NA	LINC02210- CRHR1
rs62053959	chr17_43751246_C_T	PSP	16	-1.32300	1.02E-12	ZBTB33	NA	NA	NA	LINC02210- CRHR1
rs62055663	chr17_43753454_A_G	PSP	16	-0.72733	1.62E-08	ARID3A	NA	NA	NA	LINC02210- CRHR1
rs62055708	chr17_43761974_C_A	PSP	16	0.42275	9.07E-04	JUN, JUN_FOS	NA	NA	NA	LINC02210- CRHR1
rs62055717	chr17_43766415_A_G	PSP	16	-0.53268	4.73E-05	NA	NA	NEUN	NA	LINC02210- CRHR1

rs74765707	chr17_43770416_T_C	PSP	16	0.37880	3.05E-03	NA	NA	NEUN	NA	LINC02210-CRHR1
rs17761985	chr17_43776787_C_T	PSP	16	-0.78161	9.01E-13	NA	NA	NEUN	NA	LINC02210-CRHR1
rs1623865	chr17_43778768_C_T	PSP	16	-0.38587	3.58E-03	NA	NA	NEUN	NA	LINC02210-CRHR1
rs112123127	chr17_43785098_C_T	PSP	16	0.31658	2.52E-04	NA	BRCA1, FUBP1, HNF4A	NA	NA	LINC02210-CRHR1
rs1880749	chr17_43792586_A_C	PSP	16	-0.49724	1.27E-08	ZFX	MBD2	NA	NA	LINC02210-CRHR1
rs1105570	chr17_43793342_A_G	PSP	16	-0.57392	4.00E-07	NA	RBPJ, RFX3	NA	NA	LINC02210-CRHR1
rs111511018	chr17_43794828_T_C	PSP	16	0.62701	1.57E-05	SP2, SP1	CTCF	NA	NA	LINC02210-CRHR1
rs2864087	chr17_43807063_C_T	PSP	16	-0.15195	7.32E-03	USF1, USF2, ARNT	ARNTL, BHLHE40, MAX, MTF1, MYC, MYCN, TFE3, TFEB, USF1, USF2	NA	NA	LINC02210-CRHR1
rs4609899	chr17_43807113_C_T	PSP	16	-2.01484	1.22E-39	GABPA, ELK4, ELF1	EHF, ELF1, ELF2, ELF5, ELK1, ELK3, ELK4, ERG, ETS1, ETS2, ETV7, FEV, GABPA, GABPB1+GABPB2	NA	NA	LINC02210-CRHR1
rs4471726	chr17_43807134_T_C	PSP	16	-0.65951	1.06E-15	NR2E3	NA	NA	NA	LINC02210-CRHR1
rs113790915	chr17_43810782_G_C	PSP	16	1.29869	3.26E-28	PLAG1, SP1, EGR1, SP2, EBF1, KLF5	EGR1, EGR4, KLF1, KLF3, SP1, SP1, SP3, ZIC1, ZNF148	NA	NA	LINC02210-CRHR1
rs4569335	chr17_43811950_A_C	PSP	16	0.35742	4.80E-03	NA	NA	NA	NA	LINC02210-CRHR1
rs1526126	chr17_43812218_C_T	PSP	16	-1.57052	2.60E-25	NA	NA	NA	NA	LINC02210-CRHR1
rs56006248	chr17_43815141_A_G	PSP	16	-0.82443	6.01E-08	NFIC	NA	NA	NA	LINC02210-CRHR1
rs56298110	chr17_43815336_A_G	PSP	16	0.84680	5.34E-12	NA	NA	NA	NA	LINC02210-CRHR1
rs62054424	chr17_43819384_G_A	PSP	16	-0.36192	6.28E-05	SP2, ZNF263, SP1	MAZ, SP1, SP1, SP2	NEUN	NA	LINC02210-CRHR1
rs2004260	chr17_43822398_T_C	PSP	16	0.52077	1.84E-11	NA	NA	NA	NA	LINC02210-CRHR1
rs12948216	chr17_43836673_C_T	PSP	16	-0.23658	9.67E-03	NA	ATF6	NA	NA	LINC02210-CRHR1
rs78011262	chr17_43837917_T_C	PSP	16	1.09958	1.07E-36	NFE2L1_MAF G	ATF1, ATF2+ATF4, CREB1, E4F1, HLF	NA	NA	LINC02210-CRHR1
rs28439278	chr17_43838071_C_T	PSP	16	-0.24718	2.87E-03	KLF1, KLF4	KLF1, KLF4, KLF8	NA	NA	LINC02210-CRHR1
rs80095037	chr17_43838237_G_T	PSP	16	0.24549	2.82E-03	NA	NA	NA	NA	LINC02210-CRHR1
rs62055887	chr17_43838710_T_C	PSP	16	0.46428	1.24E-07	NA	NA	NA	NA	LINC02210-CRHR1
rs62055888	chr17_43838720_C_T	PSP	16	-0.27930	1.87E-07	NA	NA	NA	NA	LINC02210-CRHR1
rs75257002	chr17_43843395_C_T	PSP	16	1.04775	1.02E-12	ESR1	FOS, FOX11, HOXB6, PBX2, PBX3	NA	NA	LINC02210-CRHR1

rs11376299 4	chr17_43847095_C_T	PSP	16	-0.22116	7.89E-03	NA	CTCF	NA	NA	LINC02210- CRHR1
rs12952301	chr17_43852487_T_C	PSP	16	0.20099	7.82E-06	NA	NA	NA	NA	LINC02210- CRHR1, CRHR1
rs11387118 1	chr17_43852733_G_A	PSP	16	0.21350	7.95E-03	NA	MYF6, MYOG, TCF12, TFAP4	NA	NA	LINC02210- CRHR1, CRHR1
rs62057068	chr17_43858629_A_G	PSP	16	1.06294	4.88E-17	NA	ETV4, SPI1	NA	NA	LINC02210- CRHR1, CRHR1
rs12940065	chr17_43865871_G_A	PSP	16	0.47643	1.56E-09	NA	NA	NA	NA	NA
rs11764650 3	chr17_43893260_T_C	PSP	16	-0.69248	7.91E-09	ARNT	ARNTL, BHLHE40, CREB1, E4F1, MAX, TFE3, TFEB, USF1, USF2	NA	NA	LINC02210- CRHR1, CRHR1
rs2316763	chr17_43895530_C_T	PSP	16	-0.51088	8.63E-14	NA	NA	NA	NA	LINC02210- CRHR1, CRHR1
rs62057108	chr17_43896616_T_C	PSP	16	-0.31100	8.54E-03	NA	NA	NEUN	NA	LINC02210- CRHR1, CRHR1
rs878887	chr17_43912582_C_T	PSP	16	-0.97194	1.68E-26	NA	NA	NA	NA	LINC02210- CRHR1, CRHR1, MAPT-AS1, SPPL2C
rs75104593	chr17_43913557_T_G	PSP	16	-1.28289	9.96E-49	NFYA, NFYB	FOXI1, NFYA, NFYB, NFYC, PBX2, PBX3	NA	NA	LINC02210- CRHR1, CRHR1, MAPT-AS1, SPPL2C
rs74998289	chr17_43913558_T_G	PSP	16	-1.78702	1.60E-42	NFYA, NFYB	FOS, FOXI1, NFYA, NFYB, NFYC, PBX2, PBX3	NA	NA	LINC02210- CRHR1, CRHR1, MAPT-AS1, SPPL2C
rs62054809	chr17_43918239_C_T	PSP	16	-0.93644	2.31E-21	NA	EOMES	NA	NA	LINC02210- CRHR1, CRHR1, MAPT-AS1, SPPL2C
rs242944	chr17_43923180_G_A	PSP	16	-0.93408	2.71E-27	NA	MYOD1, NEUROD1	NA	NA	LINC02210- CRHR1, CRHR1, MAPT-AS1, SPPL2C
rs62054825	chr17_43930033_G_A	PSP	16	-0.70869	2.63E-10	GABPA, ELK4, ELF1	EHF, ELF1, ELF2, ELK1, ELK3, ELK4, ERG, ETS1, ETS2, ETV7, FLI1, GABPA, GABPB1+GABPB2	OLIG2	NA	MAPT-AS1, SPPL2C
rs11196222 5	chr17_43934016_G_C	PSP	16	0.21365	8.12E-03	NA	NA	NA	NA	MAPT-AS1, SPPL2C
rs62054844	chr17_43937333_T_G	PSP	16	-0.36592	7.06E-06	HNF4G, HNF4A, RXRA, NR1H3_RXRA, PPARG_RXRA	ESRRA, ESRRB, NR1H2, NR2C1, NR2C2, NR2F1, NR2F1, NR2F2, NR2F6, NR4A1, NR4A2, NR4A3, PPARA, PPARG, RARG, RORA, RORC, RXRG, TCF12	NA	NA	MAPT-AS1
rs56026232	chr17_43939399_A_C	PSP	16	0.41821	2.87E-04	NA	ELF3, ERG, ETS1, FEV	NA	NA	MAPT-AS1
rs7220421	chr17_43945726_G_A	PSP	16	-0.76280	3.40E-16	KLF5	KLF1, KLF3	OLIG2	NA	MAPT-AS1
rs62055488	chr17_43950394_C_T	PSP	16	2.92186	2.58E-44	NA	NA	NA	NA	MAPT-AS1

rs80348388	chr17_43952634_A_T	PSP	16	0.42058	7.92E-03	NA	FOXA2, FOXD1, MECOM, NKX3-1	NA	NA	MAPT-ASI
rs62055496	chr17_43953307_C_T	PSP	16	0.33624	1.45E-03	NA	NA	NA	NA	MAPT-ASI
rs11242542 6	chr17_43963416_C_G	PSP	16	0.43569	1.37E-06	NA	NA	NA	NA	MAPT-ASI, MAPT, MAPT-IT1
rs62055548	chr17_43964899_A_G	PSP	16	0.19349	7.02E-03	NA	NA	NA	NA	MAPT-ASI, MAPT, MAPT-IT1
rs62055552	chr17_43965955_A_G	PSP	16	0.28282	1.06E-03	ARNT	BHLHE40, MAX, MYC, MYCN, TFEB, USF1	NA	NA	MAPT-ASI, MAPT, MAPT-IT1
rs17770296	chr17_43969136_C_T	PSP	16	-0.50129	2.05E-09	NA	NA	NA	NA	MAPT-ASI, MAPT, MAPT-IT1
rs76594404	chr17_43971457_G_C	PSP	16	0.16407	6.54E-03	MYC_MAX, MYCN	EGR1, GLI1, USF1	NA	LHX2, NEUN, OLIG2, PU1	MAPT-ASI, MAPT, MAPT-IT1
rs76324150	chr17_43973233_C_T	PSP	16	0.54586	7.30E-12	ZFX, KLF1, NFIC	ZFX	NA	LHX2, NEUN, OLIG2, PU1	MAPT-ASI, MAPT, MAPT-IT1
rs80346216	chr17_43974476_G_T	PSP	16	0.78994	2.23E-14	NA	NA	NA	LHX2, NEUN, OLIG2, PU1	MAPT-ASI, MAPT, MAPT-IT1
rs1984937	chr17_43980530_A_C	PSP	16	0.24233	5.03E-05	NA	NFAT5	LHX2, NEUN, OLIG2	NA	MAPT-ASI, MAPT, MAPT-IT1
rs4371195	chr17_43986420_A_G	PSP	16	0.41520	8.62E-16	ZNF354C	NA	NEUN	NA	MAPT
rs28439143	chr17_43987723_G_A	PSP	16	0.28427	4.47E-03	NA	STAT1, STAT3	NEUN	NA	MAPT
rs62056838	chr17_43989087_A_G	PSP	16	0.35153	3.17E-05	NA	NA	NEUN	NA	MAPT
rs62061705	chr17_44006252_T_C	PSP	16	0.35360	1.89E-09	NA	NA	NA	NA	MAPT
rs17650335	chr17_44012343_G_A	PSP	16	-0.51523	1.16E-05	PRRX2	HLF, NFIL3, VSX2, HLF, NFIL3, HLF, HNF1A, LHX3, NFIL3	NEUN, OLIG2	NA	MAPT
rs17650417	chr17_44013103_G_T	PSP	16	0.45454	1.49E-05	NA	NR2C2, YBX1	LHX2, NEUN, OLIG2	NA	MAPT
rs62061720	chr17_44013189_C_T	PSP	16	0.47477	2.89E-06	NA	IRF7, IRF8, IRF9	LHX2, NEUN, OLIG2	NA	MAPT
rs78834738	chr17_44022434_G_A	PSP	16	-0.33518	2.59E-04	NA	CBFB, KLF1, RUNX2, TBX2	LHX2, OLIG2	NA	MAPT
rs62062785	chr17_44024562_G_T	PSP	16	-0.22994	6.05E-03	NA	SMAD3, SMAD4	LHX2, OLIG2	NA	MAPT
rs62062792	chr17_44025347_T_C	PSP	16	-0.44201	1.58E-06	ARID3A	NANOG, POU4F2, T	LHX2, OLIG2	NA	MAPT
rs62062796	chr17_44027316_G_A	PSP	16	-0.27923	7.26E-07	EBF1	EBF1, TFAP2C	NEUN, OLIG2	NA	MAPT
rs76839282	chr17_44029178_G_C	PSP	16	1.34828	2.85E-11	SP1	SP1	NEUN, OLIG2	NA	MAPT
rs55946323	chr17_44031946_A_G	PSP	16	0.63539	2.17E-07	NA	NA	NEUN	NA	MAPT
rs17650860	chr17_44039008_G_A	PSP	16	-0.46047	9.85E-05	TP63, NFE2L1_MAF G	PAX2, TP53	NA	NA	MAPT
rs56280951	chr17_44041107_G_A	PSP	16	-0.39636	2.59E-06	EGR1, KLF4, KLF1, EGR2	EGR2, EGR3, GLI3, KLF4, SP1, SP1, SP3, WT1	NEUN	NA	MAPT

rs77566074	chr17_44048936_C_A	PSP	16	1.10627	2.17E-20	NA	NHLH1, POU2F1	NA	NA	MAPT
rs62063775	chr17_44050823_G_T	PSP	16	-0.22404	9.50E-03	NA	NA	NA	NA	MAPT
rs919464	chr17_44065901_C_T	PSP	16	-0.74980	5.24E-10	NA	MYOD1, NEUROD1, TCF4	NEUN	NA	MAPT
rs11238557 2	chr17_44066172_A_G	PSP	16	0.24043	5.83E-03	NA	ZBTB7B	NA	NA	MAPT
rs17651857	chr17_44069687_A_G	PSP	16	0.56076	1.29E-11	NA	NA	NA	NA	MAPT, STH
rs74759276	chr17_44069768_A_G	PSP	16	-1.51210	4.11E-23	CREB1	ATF1, ATF2+ATF4, CREM, E4F1	NA	NA	MAPT, STH
rs17651887	chr17_44069840_G_A	PSP	16	0.33414	6.85E-07	NA	NA	NA	NA	MAPT, STH
rs62063842	chr17_44070111_G_C	PSP	16	-0.44792	7.26E-07	PDX1	VSX2	NA	NA	MAPT, STH
rs62063850	chr17_44073739_G_A	PSP	16	0.42242	1.21E-18	ZNF263	NA	NEUN	NA	MAPT, STH
rs62063851	chr17_44074435_C_G	PSP	16	0.67145	6.82E-18	NA	ESR1, ESR2, MAF, MAFB, NR2C1, NR2F1, NR2F1, NR2F2, PPARG, RARG, TBX2, ZNF423	NEUN	NA	MAPT, STH
rs62064661	chr17_44078892_G_A	PSP	16	-0.20647	8.38E-03	NA	SP4	OLIG2	NA	MAPT, STH
rs62064663	chr17_44080039_T_G	PSP	16	0.19778	8.10E-04	SP2, SP1, EGR1, TFAP2C, EGR2, KLF5	EGR4, TFAP2A, TFAP2B, TFAP2C, WT1, ZNF148	OLIG2	NA	MAPT, STH
rs62064672	chr17_44086521_G_A	PSP	16	0.49949	4.24E-04	NA	ATF2+ATF4, CREM, TEF	NA	NA	MAPT, STH
rs37055887 2	chr17_44086726_T_C	PSP	16	0.46885	1.47E-03	NA	NA	NA	NA	MAPT, STH
rs62064674	chr17_44088772_C_A	PSP	16	0.18838	8.50E-03	NA	ALX1, FOXJ3	NA	NA	MAPT
rs12150260	chr17_44090538_T_G	PSP	16	0.45773	3.12E-09	NA	KLF4, SP1	NEUN	NA	MAPT
rs62062273	chr17_44092243_C_T	PSP	16	0.23081	8.04E-05	NA	NA	NA	NA	MAPT
rs62062274	chr17_44093319_G_A	PSP	16	-0.46410	5.80E-03	NA	NA	NA	NA	MAPT
rs2316948	chr17_44094667_C_G	PSP	16	0.35446	2.60E-03	NA	NA	NA	NA	MAPT
rs62062283	chr17_44095055_C_A	PSP	16	0.57880	5.82E-04	NA	ATF2+ATF4, NR2F2, RARB, RXRB, THRA, THRB, THRB	NA	NA	MAPT
rs58879558	chr17_44095467_T_C	PSP	16	0.76008	8.13E-11	NA	ATF6	NA	NA	MAPT
rs62062297	chr17_44101004_A_G	PSP	16	0.31272	1.17E-04	NA	NR6A1	LHX2, NEUN	NA	MAPT, KANSL1
rs55649944	chr17_44119463_C_T	PSP	16	-0.44515	6.12E-07	NA	NA	NA	NA	KANSL1
rs79065019	chr17_44124535_T_C	PSP	16	0.65706	1.65E-08	NA	INSM1	NA	NA	KANSL1
rs62063682	chr17_44131622_G_A	PSP	16	0.31947	6.83E-09	NA	CXXC1	NA	NA	KANSL1
rs7212857	chr17_44132301_C_A	PSP	16	-1.79011	8.62E-16	SP2, SP1	NA	NA	NA	KANSL1
rs62063685	chr17_44132379_A_G	PSP	16	-0.56746	3.10E-06	NA	BATF, ZBTB18	NA	NA	KANSL1
rs62060790	chr17_44137676_A_T	PSP	16	0.74343	6.27E-10	NA	POU2F1	NA	NA	KANSL1
rs7220752	chr17_44138201_T_C	PSP	16	0.38412	6.77E-03	NA	HOXA9, MEIS1, MEIS2	NA	NA	KANSL1
rs2316952	chr17_44139190_G_C	PSP	16	-2.59632	8.48E-11	NA	NA	NA	NA	KANSL1
rs62060847	chr17_44154946_C_T	PSP	16	0.32189	9.94E-03	NA	NR2E3	NA	NA	KANSL1
rs11236492 0	chr17_44161360_A_T	PSP	16	0.29041	8.13E-03	NA	NA	NA	NA	KANSL1
rs62061770	chr17_44169377_G_C	PSP	16	0.48839	4.28E-07	NA	NA	NA	NA	KANSL1
rs62061800	chr17_44174109_G_C	PSP	16	-0.89683	3.37E-34	NA	NA	NA	NA	KANSL1

rs62061809	chr17_44178839_T_C	PSP	16	0.28974	7.02E-03	NA	NEUROD1, SMAD2, TFAP4	LHX2, OLIG2	NA	KANSL1
rs62061813	chr17_44181118_G_A	PSP	16	-0.49128	3.28E-04	NA	NA	OLIG2	NA	KANSL1
rs4510068	chr17_44184828_G_T	PSP	16	-0.49839	4.52E-09	NA	TBX2	NA	NA	KANSL1
rs11172438 9	chr17_44209364_G_A	PSP	16	0.67594	4.14E-10	TP63	HEY2	NA	NA	KANSL1
rs75931675	chr17_44216817_A_T	PSP	16	-0.80667	2.46E-04	NA	TFE3	PU1	NA	KANSL1
rs2532314	chr17_44217112_G_T	PSP	16	0.48042	3.26E-03	NA	NA	PU1	NA	KANSL1
rs1406070	chr17_44225802_C_T	PSP	16	-0.83406	7.62E-04	NA	NA	NA	NA	KANSL1
rs2532304	chr17_44226126_A_C	PSP	16	0.42440	2.20E-06	NA	ARID5B, MEF2A, MEF2C, MEF2D	NA	NA	KANSL1
rs2532296	chr17_44231827_T_C	PSP	16	-0.31755	3.12E-03	NA	NA	NA	NA	KANSL1
rs62060864	chr17_44235893_C_T	PSP	16	0.83450	3.74E-22	NA	NA	NA	NA	KANSL1
rs75877285	chr17_44240850_T_C	PSP	16	-0.41774	2.25E-06	NA	ELF1, ELF2, ELF3, ETS1, ETS2, ETV4, ETV7, FEV	NA	NA	KANSL1
rs2732644	chr17_44252098_C_T	PSP	16	-0.36339	6.08E-05	NA	NA	NA	NA	KANSL1
rs62060917	chr17_44254467_A_G	PSP	16	0.29171	4.29E-03	NA	NA	NA	NA	KANSL1
rs2532246	chr17_44262418_G_A	PSP	16	0.27480	5.71E-04	NA	ZFX	NA	NA	KANSL1, KANSL1-AS1
rs62060921	chr17_44263479_T_C	PSP	16	0.30896	8.29E-04	NA	HEY2	NA	NA	KANSL1, KANSL1-AS1
rs76219562	chr17_44266692_C_A	PSP	16	-0.33699	8.27E-04	NA	MAZ, SP1, SP3	NA	NA	KANSL1, KANSL1-AS1
rs2732593	chr17_44267164_C_T	PSP	16	0.26892	2.81E-03	NA	NA	NA	NA	KANSL1, KANSL1-AS1
rs2696438	chr17_44272928_C_A	PSP	16	0.47721	1.18E-08	NA	MYBL2, NR6A1	NA	NA	KANSL1, KANSL1-AS1
rs2668660	chr17_44280523_A_G	PSP	16	-0.71448	5.27E-18	NA	E2F1	NA	NA	KANSL1, KANSL1-AS1
rs4792837	chr17_44286384_C_A	PSP	16	0.41144	7.02E-03	FOXP1, IRF1, PAX4	NA	NA	NA	NA
rs2907488	chr17_44302752_G_A	PSP	16	-0.29551	1.77E-10	NA	CTCF	NA	LHX2, NEUN, OLIG2, PU1	KANSL1
rs2532404	chr17_44302881_C_T	PSP	16	0.48703	7.40E-15	NA	IRF4	NA	LHX2, NEUN, OLIG2, PU1	KANSL1
rs2532403	chr17_44302904_T_C	PSP	16	0.37414	8.76E-03	NA	NA	NA	LHX2, NEUN, OLIG2, PU1	KANSL1
rs2261497	chr17_44304884_C_T	PSP	16	0.23402	1.88E-03	NA	BATF	NA	NA	KANSL1
rs14899294 8	chr17_44304910_G_C	PSP	16	-0.54788	3.63E-09	NA	NA	NA	NA	NA
rs56858903 1	chr17_44311099_G_A	PSP	16	-0.45032	9.81E-04	NA	NA	NA	NA	KANSL1
rs12603985	chr17_44313900_A_G	PSP	16	-0.65050	1.20E-05	NA	NA	NA	NA	NA
rs1706725	chr17_44324185_A_G	PSP	16	0.53869	1.29E-11	NA	ATF2+ATF4, NR2F1, NR2F2, RAR, RARG, RXRB, THRA, THRA, THRB, THRR	NA	NA	NA
rs11655181	chr17_44332093_G_A	PSP	16	-0.51940	3.41E-06	NA	SP1	NA	NA	NA
rs2696429	chr17_44335274_G_A	PSP	16	-1.11177	2.59E-07	NA	ELF2, IKZF1, SPIB	NA	NA	NA

rs78903072	chr17_44338966_A_C	PSP	16	0.62651	3.41E-06	NA	NA	NA	NA	NA
rs2732652	chr17_44345114_C_T	PSP	16	-0.45595	7.33E-04	NA	NA	LHX2, NEUN, OLIG2, PU1	NA	ARL17B
rs2458216	chr17_44346962_C_T	PSP	16	0.40067	1.54E-03	NFIC	NA	NA	NA	ARL17B
rs2696550	chr17_44349290_C_G	PSP	16	0.41061	1.74E-03	NA	NA	NA	NA	ARL17B
rs2696537	chr17_44352621_A_G	PSP	16	-0.45380	1.82E-03	NA	FOXI1	NA	NA	ARL17B
rs36039990	chr17_44352639_G_A	PSP	16	0.49710	6.51E-04	NA	NR1D1, RORA, THRA, THRB, THRB	NA	NA	ARL17B
rs113677423	chr17_44355750_C_T	PSP	16	-0.19937	4.00E-04	NA	NA	NA	NA	ARL17B
rs113896715	chr17_44356746_T_C	PSP	16	0.49946	2.17E-07	NA	NA	NA	NA	ARL17B
rs17588637	chr17_44357297_G_T	PSP	16	-0.48410	1.70E-07	NA	SOX17, SOX2	NA	NA	ARL17B
rs2696519	chr17_44357777_C_T	PSP	16	-0.42764	2.27E-05	NA	NA	NA	NA	ARL17B
rs184837332	chr17_44359783_G_T	PSP	16	-0.23429	9.15E-03	FOXD3, FOXPI	ARID3A	NA	NA	ARL17B
rs2089713	chr17_44362552_C_T	PSP	16	0.23475	7.95E-03	NA	TCF4, ZBTB18	NA	NA	ARL17B, LRRC37A
rs2696504	chr17_44362633_G_T	PSP	16	0.27598	1.34E-03	NA	NA	NA	NA	ARL17B, LRRC37A
rs2732614	chr17_44364209_C_T	PSP	16	-0.70554	4.67E-11	NA	CTCF	NA	NA	ARL17B, LRRC37A
rs2732611	chr17_44365487_A_C	PSP	16	-0.39973	6.09E-04	NA	HOXC6	NA	NA	ARL17B, LRRC37A
rs2668650	chr17_44367939_G_A	PSP	16	-0.26637	8.27E-04	NA	NA	NA	NA	ARL17B, LRRC37A
rs183211	chr17_44788310_G_A	PSP	16	0.83182	2.24E-03	NA	NA	NEUN	NA	NSF
rs9912530	chr17_44836302_T_C	PSP	16	-1.76031	9.32E-15	PAX5	NHLH1	NEUN	NA	NSF, WNT3
rs141740390	chr17_61528837_T_C	AD	17	0.32205	1.52E-03	NA	NA	NA	NA	CYB561
rs57047618	chr18_29081760_C_G	AD	18	0.85742	5.85E-05	SRF, NR4A2	ESR2, ESRRB, ESRRG, NR1D1, NR2F1, NR2F2, NR4A1, NR4A2, NR4A3, PPARA, RORA, RORC, RXRB, SRF, THRA, THRB, THRB	NA	NA	DSG2
rs59665004	chr18_29082231_T_C	AD	18	0.47126	4.60E-04	NA	RUNX1	NA	NA	DSG2
rs2927437	chr19_45241638_A_G	AD	20	0.46793	7.40E-08	NA	MAZ, TP73	NA	NA	NA
rs8100197	chr19_45253582_G_A	AD	20	-0.19227	1.94E-03	NA	FOXO1, FOXO3, SRY	NA	NA	BCL3, MIR8085
rs2965113	chr19_45301924_T_C	AD	20	1.82116	4.01E-19	SP2, CTCF	ETV7, GABPA, GABPB1+GABPB2	NA	NA	CBLC
rs10412413	chr19_45327309_C_T	AD	20	-0.57248	1.08E-05	NA	RUNX3, TLX1	NA	NA	BCAM
rs11667241	chr19_45337504_G_C	AD	20	0.32384	3.51E-04	PAX5, SPIB	NA	NA	NA	NA
rs10407439	chr19_45341948_A_G	AD	20	0.49827	5.04E-07	NA	TCF4	NA	NA	NECTIN2
rs10408472	chr19_45342124_G_A	AD	20	-0.36309	2.31E-06	SMAD2, SMAD3, SMAD4	TBX2, TBX3, TBX5, TCF4	NA	NA	NECTIN2
rs2972562	chr19_45343664_A_G	AD	20	0.39260	4.59E-03	NR2C2, SP2, SP1	NR2F2, PPARA, RXRA, SP1, SP1, SP2, SP3, WT1	NA	NA	NECTIN2

rs2199575	chr19_45345623_G_A	AD	20	-0.31984	8.91E-06	NA	ESRRA, NR2C1, RARB, RARG, RXRB	NA	NA	NECTIN2
rs2376518	chr19_45345773_C_T	AD	20	-0.47736	8.48E-11	SP2, SP1	NA	NA	NA	NECTIN2
rs11666329	chr19_45354296_A_G	AD	20	-0.33234	3.80E-08	SP2	NA	NA	NA	NECTIN2
rs73050205	chr19_45356464_T_A	AD	20	-0.18790	5.77E-03	FOXD3	ALX1, FOXC2, IRF9, OTX2	NA	NA	NECTIN2
rs143459034	chr19_45358026_C_T	AD	20	-0.32260	6.02E-03	NA	CEBPG	NA	NA	NECTIN2
rs555608	chr19_45365072_A_G	AD	20	0.36089	3.41E-07	NA	YY1	NA	NA	NECTIN2
rs3112439	chr19_45367972_G_C	AD	20	0.32192	4.32E-09	NA	NA	NA	NA	NECTIN2
rs565566	chr19_45370570_A_C	AD	20	0.47063	1.35E-04	NA	NRF1	NA	NA	NECTIN2
rs11672399	chr19_45377098_C_A	AD	20	-0.40199	2.81E-03	NA	NA	NA	NA	NECTIN2
rs34278513	chr19_45378144_C_T	AD	20	-0.37890	7.34E-08	NA	NA	NA	NA	NECTIN2
rs387976	chr19_45379060_A_C	AD	20	-0.22770	4.65E-03	NA	NA	NA	NA	NECTIN2
rs412776	chr19_45379516_G_A	AD	20	1.04686	7.92E-10	NA	GLI3	NA	NA	NECTIN2
rs12972156	chr19_45387459_C_G	AD	20	-0.42130	2.35E-04	TFAP2C	TFAP2A, TFAP2B, TFAP2C	NA	NA	NECTIN2, TOMM40
rs115881343	chr19_45403216_C_T	AD	20	-0.35783	1.04E-04	ZNF263, SP1, SP2	NA	NA	NA	TOMM40, APOE
rs1160985	chr19_45403412_C_T	AD	20	-0.35579	7.62E-04	NA	NA	NA	NA	TOMM40, APOE
rs14162290	chr19_45426792_G_A	AD	20	-0.61399	6.66E-09	NA	NA	LHX2, PU1	NA	APOC1, APOC1P1
rs12977604	chr19_45442528_C_G	AD	20	-0.28871	6.06E-03	NA	BACH1, ZNF143	NA	NA	APOC1P1, APOC4, APOC4-APOC2, APOC2
rs73045696	chr19_45443088_G_C	AD	20	-0.49902	1.34E-05	NA	BCL6, ZBTB33	NA	NA	APOC1P1, APOC4, APOC4-APOC2, APOC2
rs10423208	chr19_45453656_G_A	AD	20	-0.24192	3.14E-04	NA	HNF4A, NR2F6	LHX2, OLIG2, PU1	NA	APOC4, APOC4-APOC2, APOC2, CLPTM1
rs117316645	chr19_45458212_G_A	AD	20	-0.92965	1.31E-34	NA	NA	NA	LHX2, NEUN, OLIG2, PU1	APOC4, APOC4-APOC2, APOC2, CLPTM1
rs8100120	chr19_45466058_T_C	AD	20	0.84827	5.27E-20	SMAD2_SMA D3_SMAD4	TBX2	NA	NA	CLPTM1
rs16979600	chr19_45487420_A_C	AD	20	0.36534	6.46E-08	NA	NA	NA	NA	CLPTM1
rs2004357	chr19_45618959_G_A	AD	20	-0.47191	1.58E-06	NA	NRF1, NRF1	NA	NA	PPP1R37
rs28620490	chr19_45638245_C_T	AD	20	-0.22542	7.79E-06	NA	HEY2	NEUN	NA	PPP1R37
rs10417602	chr19_45660639_C_T	AD	20	0.40362	7.02E-03	ZNF263	IRF5	NA	NA	NKPD1, TRAPPC6A
rs139330969	chr19_45712550_G_A	AD	20	-0.52030	4.63E-08	NA	GLI3	NA	NA	EXOC3L2
rs182645183	chr19_45712970_T_C	AD	20	0.22826	5.34E-04	NA	KLF4, KLF8, ZNF143	NA	NA	EXOC3L2
rs56845483	chr19_45713298_T_C	AD	20	-0.42353	4.42E-03	NA	FOXC2	NA	NA	EXOC3L2
rs112759099	chr19_45726845_C_G	AD	20	-0.54989	5.50E-04	NR1H3_RXRA	EGR3, ESR1, ZNF219	NA	NA	EXOC3L2
rs3845674	chr2_127875389_T_G	AD	22	-0.43915	2.88E-03	NA	TCF4	OLIG2	NA	NA
rs4663099	chr2_127878432_C_A	AD	22	0.30010	1.88E-05	NA	ZBTB4	OLIG2	NA	NA
rs10929009	chr2_127878839_T_C	AD	22	-0.40690	8.00E-04	NFE2L1_MAF G	RARB, RARG, RXRB, THR B	OLIG2	NA	NA
rs13025717	chr2_127886158_C_T	AD	22	-1.77075	2.60E-38	KLF4, KLF5, KLF1	KLF4, SP1, SP1, SP2, SP3	OLIG2, PU1	NA	NA

rs13004848	chr2_127887560_T_C	AD	22	-0.27923	1.50E-04	NA	NA	OLIG2	NA	NA
rs6710467	chr2_127889932_G_A	AD	22	-0.61937	8.25E-04	NA	NA	PU1	NA	NA
rs56368748	chr2_127894098_G_A	AD	22	0.65095	4.32E-08	NA	SMAD3, SMAD4	OLIG2, PU1	NA	NA
rs36181881	chr2_234055276_A_G	AD	23	0.26719	1.11E-04	NA	NA	NA	NA	INPP5D
rs75322627	chr2_234068192_T_A	AD	23	0.25967	2.36E-03	NA	MEF2D	NA	NA	INPP5D
rs6024861	chr20_54984140_G_A	AD	24	-0.11299	2.81E-03	NA	KLF3	PU1	NA	CSTF1, CASS4
rs6064392	chr20_54984768_G_T	AD	24	-1.02889	2.42E-20	NA	ATF6, ATF6	PU1	NA	CSTF1, CASS4
rs7274581	chr20_55018260_T_C	AD	24	0.29915	1.74E-03	NA	NA	PU1	NA	CASS4
rs16979933	chr20_55020557_T_C	AD	24	-0.17341	7.89E-03	NA	NA	NA	NA	CASS4
rs2830489	chr21_28148191_C_T	AD	25	0.15454	8.00E-03	NA	TGIF1	NA	NA	NA
rs12494100	chr3_39481835_G_A	PSP	26	-0.32193	6.09E-04	CEBPB	DDIT3	NA	NA	NA
rs57388308 3	chr3_39799389_G_A	PSP	26	-0.51794	9.03E-05	NA	TBX2, TBX5	NA	NA	NA
rs79138148	chr6_32516351_C_T	AD	27	0.33109	2.22E-03	NA	NA	NA	NA	HLA-DRB6, HLA-DRB1, NA, HLA-DRB4
rs11173930 6	chr6_32516848_T_G	AD	27	-0.33195	7.79E-06	NA	NA	NA	NA	HLA-DRB6, HLA-DRB1, NA, HLA-DRB4
rs11570790 5	chr6_32516930_G_C	AD	27	0.22294	1.69E-03	NA	NA	NA	NA	HLA-DRB6, HLA-DRB1, NA, HLA-DRB4
rs11208053 6	chr6_32521594_A_T	AD	27	0.40112	3.54E-03	NA	NA	NA	NA	HLA-DRB6, HLA-DRB1, NA, HLA-DRB4
rs9269817	chr6_32549746_A_G	AD	27	-0.60472	5.45E-10	USF1, USF2, NR4A2	ESR2, ESRRB, ESRRG, NR4A2, NR4A3, PPARA, RARG, RXRB, THRA, THRB, TISE1	NA	NA	HLA-DRB1
rs9270924	chr6_32572408_G_A	AD	27	-0.44397	1.85E-03	NA	NA	PU1	NA	NA
rs9270953	chr6_32573045_C_T	AD	27	-0.77890	1.24E-07	SMAD2, SMAD3, SMAD4	NA	PU1	NA	NA
rs9270954	chr6_32573058_C_T	AD	27	0.92744	7.78E-11	NA	NA	PU1	NA	NA
rs9271127	chr6_32576887_C_T	AD	27	-0.99667	1.95E-26	NA	NA	PU1	NA	NA
rs9271129	chr6_32576889_A_T	AD	27	-0.89479	1.68E-14	NA	NA	PU1	NA	NA
rs9271162	chr6_32577715_G_T	AD	27	-0.29971	1.15E-03	NA	NA	PU1	NA	NA
rs9271164	chr6_32577784_C_T	AD	27	0.24730	6.10E-05	NA	NA	PU1	NA	NA
rs9271171	chr6_32577907_C_T	AD	27	-0.35346	4.93E-06	NA	NA	PU1	NA	NA
rs2790095	chr6_45439307_C_G	PSP	28	0.22085	1.31E-03	TCF3, NHLH1, ZNF263, MZF1_1-4	NA	NA	NA	RUNX2
rs9473122	chr6_47474962_C_T	AD	29	0.25169	5.91E-03	NA	MEF2A, MEF2C, MEF2D	NA	NA	CD2AP
rs9473123	chr6_47475339_G_A	AD	29	-0.38600	5.88E-05	KLF1, KLF4	KLF1, KLF3	NA	NA	CD2AP
rs9349416	chr6_47551938_G_A	AD	29	-0.70539	4.33E-03	SOX10	NA	NA	NA	CD2AP
rs75045569	chr7_143109208_T_G	AD	31	-0.75662	1.05E-04	NA	CBFB, RUNX1, RUNX3	NA	NA	EPHA1, EPHA1-AS1
rs62472730	chr7_143123975_T_A	AD	31	0.47792	7.41E-03	NA	NA	NA	NA	EPHA1-AS1
rs10224310	chr7_143126994_A_C	AD	31	-0.35532	3.09E-06	NA	SPI1	NA	NA	EPHA1-AS1
rs755951	chr8_27226790_A_C	AD	32a	0.76040	7.05E-08	CTCF	CTCF	NEUN, PU1	NA	PTK2B
rs17057051	chr8_27227554_A_G	AD	32a	0.47831	1.34E-04	NA	HBPI1, LHX2	NEUN	NA	PTK2B

rs1532277	chr8_27466181_T_C	AD	32b	0.26338	4.68E-05	ARNT	ARNT, ARNTL, BHLHE40, MAX, MYC, TFEB, USF1	LHX2, NEUN, OLIG2	NA	CLU, MIR6843
rs12547073	chr8_131043810_G_A	PSP	33	-0.65927	1.57E-10	NA	GABPA, GABPB1+GABPB2	NA	NA	NA
rs2045091	chr8_131075859_C_T	PSP	33	-0.24980	6.57E-04	NFE2L1_MAF G	NA	NEUN	NA	ASAP1

All 320 variants with significant allelic skew (SigVars) identified across both MPRA stages. Locus annotations are provided in Supplemental Data. SigVars are annotated with predicted TFBS-disruption using two algorithms (TFBS_1: SNPS2TFBS, TFBS_2: motifbreakR), and cell-type specific Enhancer and Promoter marks from human brain provided by Nott *et al.*, 2019. Finally, variants are annotated for cis genes (gene body +/- 10 kb).

Table S2. 17q21.31 LD clusters, SigVars, and cis-Genes

LD Block	LD Clusters	Number of SigVars	cis Genes
1	4	2	GJC1, HIGD1B, EFTUD2
2	3	3	MAP3K14-AS1, SPATA32, MAP3K14
3	2, 5	12	ARHGAP27, PLEKHM1, LRRC37A4P
4	1,6,7	151	MAPK8IP1P2 ,LINC02210, CRHR1, MAPT-AS1, SPPL2C, MAPT, STH, KANSL1, ARL17B, LRRC37A
NA	NA	20	NA

188 SigVars identified in the 17q21.31 region clustered by LD (CLQD function; Methods; Source Data for individual variant annotations) further organized into contiguous LD blocks. The number of SigVars per LD block and annotated cis genes (+/- 10kb) are shown.

Table S3. SigVars per gene

Supplemental Table 3: SigVars per gene

Genes for which there was an MPRA SigVar within +/- 10 kb of the gene body.

"SigVar Number" is the number SigVars/gene. Some SigVars are annotated to multiple nearby genes.

chr	Gene	Sigvar Number
chr15	ADAM10	1
chr11	AGBL2	1
chr12	AMHR2	4
chr19	APOC1	1
chr19	APOC1P1	3
chr19	APOC2	4
chr19	APOC4	4
chr19	APOC4-APOC2	4
chr19	APOE	2
chr17	ARHGAP27	7
chr17	ARL17B	15
chr8	ASAP1	1
chr19	BCAM	1
chr19	BCL3	1
chr11	C1QTNF4	1
chr20	CASS4	4
chr19	CBLC	1
chr6	CD2AP	3
chr11	CELF1	1
chr19	CLPTM1	4
chr8	CLU	1
chr1	CR1	2
chr17	CRHR1	11
chr20	CSTF1	2
chr17	CYB561	1
chr17	DCAKD	1
chr18	DSG2	2
chr17	EFTUD2	1
chr7	EPHA1	1
chr7	EPHA1-AS1	3
chr19	EXOC3L2	4
chr11	FAM180B	1
chr14	FERMT2	2
chr11	FNBP4	1
chr17	GJC1	1
chr17	HIGD1B	1
chr6	HLA-DRB1	5
chr6	HLA-DRB4	4
chr6	HLA-DRB6	4
chr2	INPP5D	2

chr16	IQCK	15
chr17	KANSL1	37
chr17	KANSL1-AS1	6
chr11	KBTBD4	1
chr16	KNOP1	5
chr17	LINC02210	3
chr17	LINC02210-CRHR1	45
chr1	LINC02257	2
chr15	LOC101928725	2
chr17	LRRC37A	5
chr17	LRRC37A4P	16
chr17	MAP3K14	3
chr17	MAP3K14-AS1	2
chr17	MAPK8IP1P2	8
chr17	MAPT	45
chr17	MAPT-AS1	21
chr17	MAPT-IT1	8
chr15	MINDY2	2
chr17	MIR4315-1	4
chr17	MIR4315-2	4
chr8	MIR6843	1
chr19	MIR8085	1
chr1	MR1	2
chr11	MS4A4E	1
chr11	MS4A6A	1
chr11	NDUFS3	1
chr19	NECTIN2	16
chr19	NKPD1	1
chr17	NMT1	1
chr17	NSF	2
chr17	PLEKHM1	11
chr19	PPP1R37	2
chr12	PRR13	1
chr8	PTK2B	2
chr6	RUNX2	1
chr14	SLC24A4	1
chr12	SP1	3
chr17	SPATA32	1
chr17	SPPL2C	7
chr17	STH	10
chr1	STX6	1
chr19	TOMM40	3
chr19	TRAPPC6A	1
chr16	VPS35L	3
chr17	WNT3	1

Genes for which there was an MPRA SigVar within +/- 10 kb of the gene body. "SigVar Number" is the number SigVars/gene. Some SigVars are annotated to multiple nearby genes.

Data S1: Supplemental Data

This file provides summary statistics for all variants tested in MPRA stages 1 and 2

Data S2: Source Data

This file provides raw data and computed scores used for downstream analyses and figure construction. Data are organized by relevant figure.

Data S3: Supplemental Information - Methods

This file provides additional Methods data including sequences, primers, and accessions.