ARTICLES
https://doi.org/10.1038/s41588-020-00721-x

Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases

M. Ryan Corces[1,2], Anna Shcherbina[3,4], Soumya Kundu[4,5], Michael J. Gloudemans[1,3], Laure Frésard[1], Jeffrey M. Granja[2,4,6], Bryan H. Louie[1,2], Tiffany Eulalio[1,3], Shadi Shams[2,4], S. Tansu Bagdatli[2,4], Maxwell R. Mumbach[2,4], Boxiang Liu[1,7,8], Kathleen S. Montine[1], William J. Greenleaf[2,4,9,10], Anshul Kundaje[4,5], Stephen B. Montgomery[1,4], Howard Y. Chang[2,4,11,12] ✉ and Thomas J. Montine[1] ✉

Genome-wide association studies of neurological diseases have identified thousands of variants associated with disease phenotypes. However, most of these variants do not alter coding sequences, making it difficult to assign their function. Here, we present a multi-omic epigenetic atlas of the adult human brain through profiling of single-cell chromatin accessibility landscapes and three-dimensional chromatin interactions of diverse adult brain regions across a cohort of cognitively healthy individuals. We developed a machine-learning classifier to integrate this multi-omic framework and predict dozens of functional SNPs for Alzheimer's and Parkinson's diseases, nominating target genes and cell types for previously orphaned loci from genome-wide association studies. Moreover, we dissected the complex inverted haplotype of the *MAPT* (encoding tau) Parkinson's disease risk locus, identifying putative ectopic regulatory interactions in neurons that may mediate this disease association. This work expands understanding of inherited variation and provides a roadmap for the epigenomic dissection of causal regulatory variation in disease.

Alzheimer's (AD) and Parkinson's (PD) diseases affect approximately 50 and 10 million individuals worldwide, respectively, as two of the most common neurodegenerative disorders. Several large consortia have assembled genome-wide association studies (GWAS) that associate genetic loci with clinical diagnoses of probable AD dementia[1–4] or probable PD[5–7], or with their characteristic pathological features. These efforts have led to the identification of dozens of potential risk loci for these diseases. However, most risk loci reside in noncoding regions, and so it is unclear if the nominated (often nearest) gene is functionally relevant for the disease or if another gene is involved[8].

Most functional noncoding SNPs would be predicted to exert their effects through the alteration of gene expression via perturbation of transcription factor binding and regulatory element function[8]. Such regulatory elements are highly cell type specific[9], suggesting that the resultant effects of noncoding SNPs would be equally cell type specific. Thus, comprehensive nomination of putative functional noncoding SNPs in the brain requires cataloging the regulatory elements that are active in every brain cell type in the correct organismal and regional context. These critical data hold the promise to illuminate the functional importance of genetic risk loci in the molecular pathogenesis of common neurodegenerative diseases.

Previous work has carefully mapped such cell-type-specific gene regulatory landscapes in the human brain, predominantly during early developmental time points[10], in organoid culture systems[11–13] or in induced pluripotent stem cell-derived cellular models[14,15]. Additional studies have profiled chromatin accessibility in macrodissected postmortem adult human brain[16–19]. Such datasets have provided a rich resource for the nomination of putative functional SNPs in neurological disease by using multi-omic approaches[10,14,17,20]. Moreover, recent work has profiled chromatin accessibility and 3D chromatin conformation in primary brain cell types from resected pediatric brain tissue to explore the roles of noncoding SNPs in AD[9]. Lastly, innovative analytical approaches, for example, leveraging machine learning, have greatly expanded our ability to predict the functional effects of noncoding SNPs[21–25]. Cumulatively, this work has provided important advances in our understanding of the role of noncoding SNPs in disease predisposition, particularly in neurological disease.

In this study, we build on the current understanding of inherited variation in neurodegenerative disease through the implementation of a multi-omic framework that enables accurate prediction of functional noncoding SNPs. This framework layers bulk assay for transposase-accessible chromatin using sequencing (ATAC–seq)[26], single-cell ATAC–seq (scATAC–seq)[27] and HiChIP enhancer

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [2]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [5]Department of Computer Science, Stanford University, Stanford, CA, USA. [6]Program in Biophysics, Stanford University, Stanford, CA, USA. [7]Department of Biology, Stanford University, Stanford, CA, USA. [8]Baidu Research, Sunnyvale, CA, USA. [9]Department of Applied Physics, Stanford University, Stanford, CA, USA. [10]Chan Zuckerberg Biohub, San Francisco, CA, USA. [11]Program in Epithelial Biology, Stanford University, Stanford, CA, USA. [12]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ✉e-mail: howchang@stanford.edu; tmontine@stanford.edu

connectome[28,29] data over a machine-learning classifier to predict putative functional SNPs driving the association with neurodegenerative diseases. Through these efforts, we pinpoint putative target genes and cell types of several noncoding GWAS loci in AD and PD, providing a roadmap for the application of these data and technology to other neurological disorders and enabling a more comprehensive understanding of the role of inherited noncoding variation in disease.

Results

Bulk chromatin accessibility landscapes in macrodissected tissue identify brain-regional epigenomic heterogeneity. We profiled the bulk chromatin accessibility landscapes of 7 macrodissected brain regions across 39 cognitively healthy individuals to characterize the role of the noncoding genome in neurodegenerative diseases (Supplementary Table 1). These brain regions include distinct isocortical regions (superior and middle temporal gyri, parietal lobe and middle frontal gyrus), striatal regions (caudate nucleus and putamen), the hippocampus and the substantia nigra (Fig. 1a and Methods). From these bulk ATAC–seq libraries, we compiled a merged set of 186,559 reproducible peaks (Fig. 1b and Supplementary Data 1). In this study, a reproducible peak is defined as any peak that is called in at least 30% of the bulk ATAC–seq samples from any given brain region (Supplementary Fig. 1a and Methods). Dimensionality reduction via *t*-distributed stochastic neighbor embedding (*t*-SNE) identified four distinct clusters of samples, grouped roughly by major brain regions (Fig. 1c). While many region-specific peaks in chromatin accessibility could be identified from these bulk ATAC–seq data, most of these peaks corresponded to cell types predominantly present in a single region (Fig. 1d). A detailed analysis of these bulk ATAC–seq data primarily revealed region-specific differences in chromatin accessibility (Supplementary Fig. 1b–h and Supplementary Note 1).

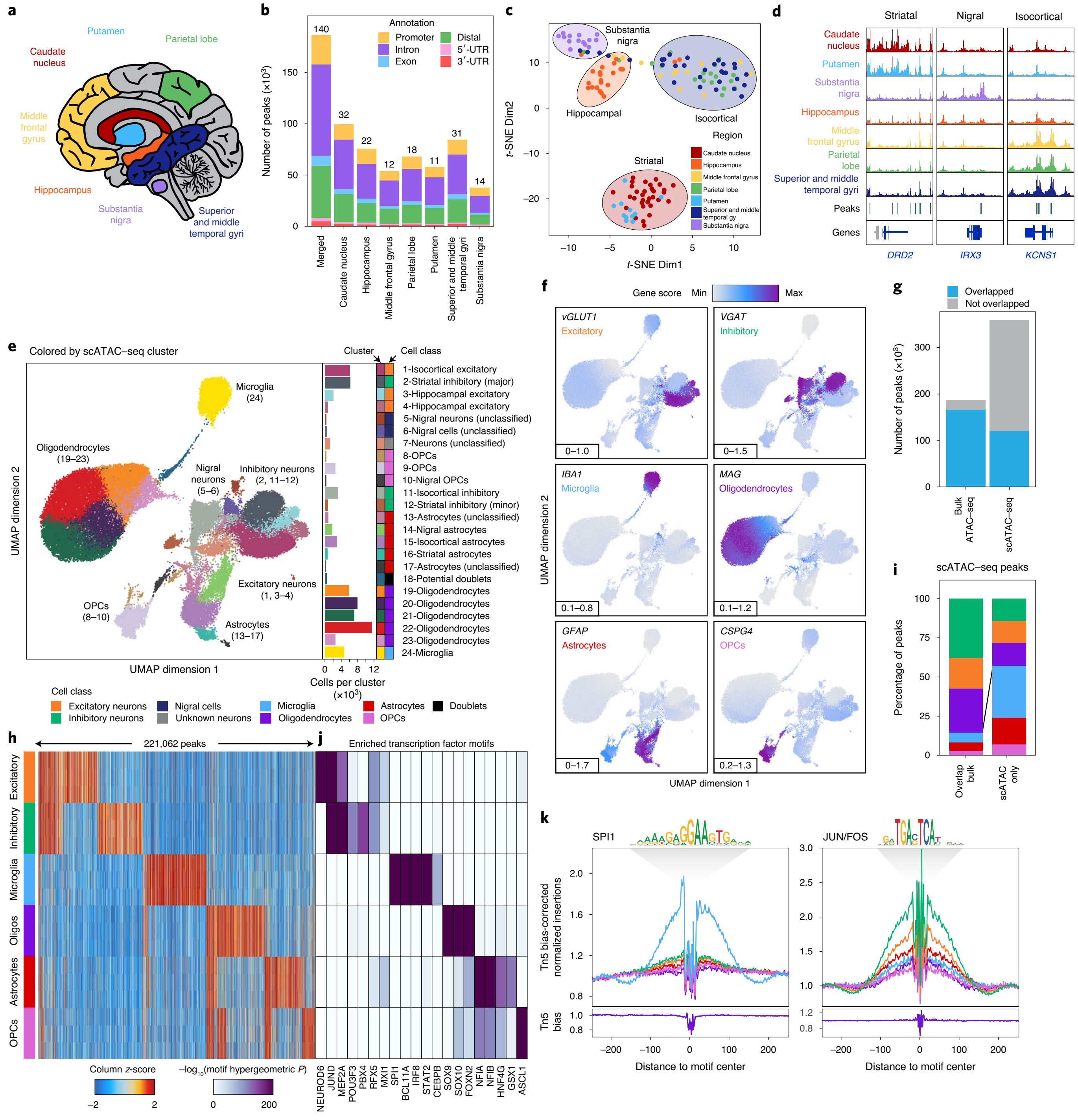
scATAC–seq captures regional and cell-type-specific heterogeneity. To better understand brain-regional cell-type-specific chromatin accessibility landscapes, we performed single-cell chromatin accessibility profiling in 10 samples spanning the isocortex ($n=3$), striatum ($n=3$), hippocampus ($n=2$) and substantia nigra ($n=2$) (Supplementary Table 1). In total, we profiled chromatin accessibility in 70,631 individual cells (Fig. 1e) after stringent quality control filtration (Supplementary Fig. 2a and Supplementary Data 2). Unbiased iterative clustering[27,30] and Harmony-based batch correction of these single cells identified 24 distinct clusters (Fig. 1e and Extended Data Fig. 1a,b), which were assigned to known brain cell types based on gene activity scores compiled from chromatin accessibility signal in the vicinity of key lineage-defining genes[30,31] (Fig. 1f, Extended Data Fig. 1c,d and Methods). Additionally, 13 of the 24 clusters showed regional specificity with some clusters composed almost entirely from a single brain region (Extended Data Fig. 1e,f and Supplementary Data 2). We did not identify any clusters that were clearly segregated by sex but the sample size used in this study was not powered to make such a determination (Extended Data Fig. 1g). Cumulatively, we defined eight distinct cell classes, including the six main brain cell types (excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, astrocytes and oligodendrocyte progenitor cells (OPCs)) and identified one cluster (cluster 18) as putative doublets that we excluded from the downstream analyses (Fig. 1e and Extended Data Fig. 1h). These cell groupings varied largely in the total number of cells per grouping (Extended Data Fig. 1i) and showed distinct donor and regional compositions (Extended Data Fig. 1j–m).

Using these clusters, we then called peaks from scATAC–seq pseudobulk chromatin accessibility to create a union set of 359,022 reproducible peaks (Supplementary Data 3). Overall, 89% of bulk ATAC–seq peaks were overlapped by a peak called in the scATAC–seq data (Fig. 1g). Conversely, only 34% of scATAC–seq peaks were overlapped by a peak from the bulk ATAC–seq peak set (Fig. 1g). Consistent with a role for distal regulatory elements in cell-type-specific gene regulation[32], we found an enrichment in distal/intronic peaks and a depletion in promoter peaks in the peak set specifically identified via scATAC–seq (Extended Data Fig. 2a). To better understand the cell type specificity of the scATAC–seq peaks, we identified cell-type-specific peaks through ‘feature binarization’, which identifies peaks that are uniquely accessible in a single cell type or subset of cell types[33]. This analysis identified 221,062 highly cell-type-specific peaks within the 6 primary brain cell types, comprising >60% of all peaks identified from our scATAC–seq data (Fig. 1h and Supplementary Data 4). These cell-type-specific peaks were also enriched for distal/intronic peaks and depleted for promoter peaks (Extended Data Fig. 2b). Some of these peaks were shared across the different neuronal cell types while others were shared across astrocytes, OPCs and oligodendrocytes (Fig. 1h, Extended Data Fig. 2c and Supplementary Data 4). However, 48% of peaks called in our scATAC–seq data were specific to a single cell type ($n=172{,}111$ peaks; Fig. 1h and Supplementary Data 4) with the vast majority of these cell-type-specific peaks remaining undetected in our bulk ATAC–seq analyses. Consistent with previous work[34], we found an enrichment of peaks from less abundant cell types (less than 20% of cells, that is, microglia, astrocytes and OPCs) within the set of peaks identified via scATAC–seq but not bulk ATAC–seq

Fig. 1 | scATAC–seq identifies cell-type-specific chromatin accessibility in the adult brain. a, Brain regions profiled in this study. **b**, Bar plot showing the number of reproducible peaks identified from samples in each brain region. The ‘Merged’ bar represents the final merged peak set. The numbers above each bar represent the total number of biological samples profiled for each brain region. **c**, *t*-SNE dimensionality reduction of bulk ATAC–seq data. Each dot represents a single piece of tissue with technical replicates merged where applicable. **d**, Sequencing tracks of region-specific ATAC–seq peaks. From left to right, *DRD2* (striatum specific; chr11:113,367,951–113,538,919), *IRX3* (substantia nigra specific; chr16:54,276,577–54,291,319) and *KCNS1* (isocortex specific; chr20:45,086,706-45,107,665). The tracks have been normalized to the total number of reads in TSS regions. **e**, Left: uniform manifold approximation and projection (UMAP) dimensionality reduction after iterative LSI of scATAC–seq data from 10 different samples. Each dot represents a single cell ($n=70{,}631$), colored by its corresponding cluster. Right: Bar plot showing the number of cells per cluster. **f**, Same as Fig. 1e but each cell is colored by its gene activity score for the annotated lineage-defining gene. The minimum and maximum gene activity scores are shown in the bottom left of each panel. **g**, Bar plot showing the overlap of bulk ATAC–seq and scATAC–seq peak calls. ‘Bulk ATAC–seq’ represents the number of peaks from the bulk ATAC–seq merged peak set that are overlapped by a peak called in our scATAC–seq merged peak set. ‘scATAC–seq’ represents the number of peaks from our scATAC–seq merged peak set that are overlapped by a peak called in our bulk ATAC–seq merged peak set. Overlap is considered as any overlapping bases. **h**, Heatmap representation of chromatin accessibility in binarized peaks ($n=221{,}062$) from the scATAC–seq peak set. Each row represents an individual pseudobulk replicate (three per cell type) and each column represents a peak. Oligos, oligodendrocytes. **i**, Bar plot of the percentage of peaks from the scATAC–seq binarized peak set that overlap peaks identified by the bulk ATAC–seq (‘Overlap bulk’) or are uniquely identified by scATAC–seq (‘scATAC only’). Only peaks found to be unique to a single cell type ($n=172{,}111$) were used in this analysis. The bars are colored according to the legend above Fig. 1h. **j**, Motif enrichments of the binarized peaks identified in Fig. 1h. Due to redundancy in motifs, transcription factor drivers were predicted using the average gene expression in GTEx brain samples and accessibility at transcription factor promoters in cell class-grouped scATAC–seq profiles. **k**, Footprinting analysis of the SPI1 (left; CIS-BP M6484_1.02) and JUN/FOS (right; CIS-BP M4625_1.02) transcription factors across the six major cell classes.

(Fig. 1i and Extended Data Fig. 1l). Similarly, examining per-cell accessibility at the peaks specifically identified via scATAC–seq, we found significantly fewer cells supporting these peaks (Extended Data Fig. 2d). These results highlight the utility of single-cell methods when cell-type-specific peaks are difficult to identify from bulk tissues containing multiple distinct cell types at varying frequencies.

To predict which transcription factors may be responsible for establishing and maintaining these cell-type-specific regulatory programs, we performed motif enrichment analyses of peaks specific to each cell type (Fig. 1j). We identified many known drivers of cell type identity, such as motifs specific to SOX9 and SOX10 in oligodendrocytes[35,36] or to ASCL1 in OPCs[37,38]. Lastly, transcription factor footprinting from our scATAC–seq-derived cell-type-specific chromatin accessibility data showed enrichment of binding of key lineage-defining transcription factors such as SPI1 in microglia[39] and JUN/FOS in neurons[40] (Fig. 1k). Notably, the three isocortical samples, derived from distinct brain regions, showed high similarity based on Pearson correlation, supporting their use as biological replicates (Extended Data Fig. 2e). These data provide reference cell profiles for cell-type-specific deconvolution of bulk ATAC–seq data (Supplementary Fig. 3, Supplementary Data 5 and Supplementary Note 2) and identify brain-regional heterogeneity in glial cells, such as astrocytes and OPCs (Supplementary Fig. 4, Supplementary Data 6 and Supplementary Note 3).



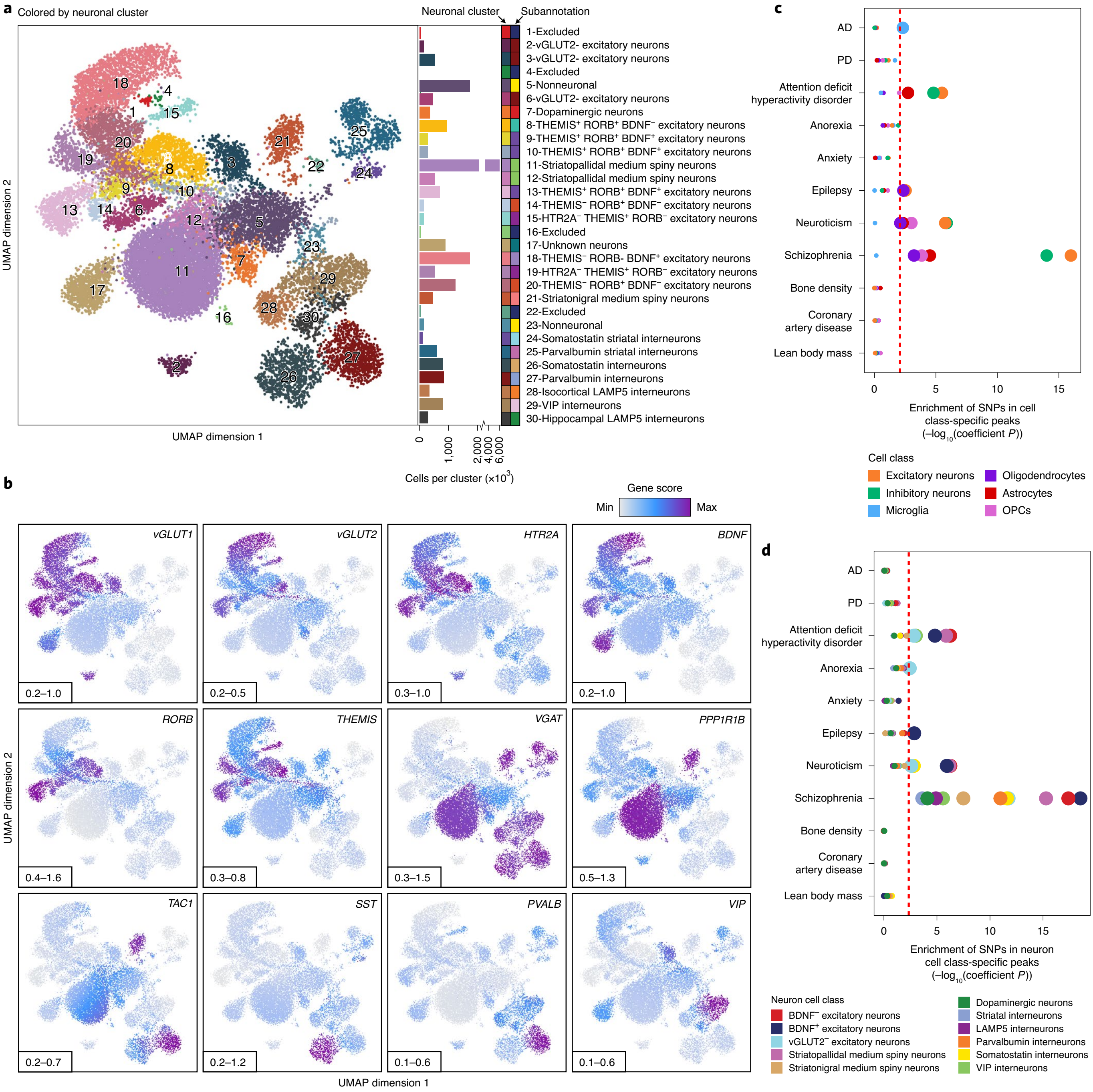


Fig. 2 | Subclustering identifies diverse biologically relevant neuronal cell types in the adult brain. a, Left: UMAP dimensionality reduction after iterative LSI of scATAC–seq data from neuronal cells from ten different samples. Each dot represents a single cell ($n=21{,}116$). The dots are colored by their corresponding neuronal subcluster. Neuronal cluster numbers are overlaid on the UMAP above each neuronal cluster centroid. Right: Bar plot showing the number of cells per cluster. Each neuronal cluster subannotation is labeled to the right of the bar plot and indicated by color. **b**, The same UMAP dimensionality reduction shown in Fig. 2a but each cell is colored by its gene activity score for the annotated lineage-defining gene. The minimum and maximum gene activity scores are shown in the bottom left of each panel. **c**,**d**, LD score regression identifying the enrichment of GWAS SNPs from various brain-related and nonbrain-related conditions in the peak regions of various cell classes from the broad scATAC–seq clustering (**c**) or neuronal cell classes identified from the neuronal subclustering analysis (**d**). The dashed line represents the Bonferroni-corrected significance threshold for the LDSC coefficient *P* value (Methods) adjusted for the number of cell classes tested. The size of the point for each cell class indicates whether this cell class passed the Bonferroni-corrected significance threshold (larger) or not (smaller).

scATAC–seq identifies diverse neuronal subpopulations. Given the well-understood diversity of neuronal types and functions, we sought to further subdivide our scATAC–seq data based on neuronal subtypes. Extracting all cells previously labeled as neurons (clusters 1–7, 11 and 12; $n=21{,}116$ cells), we performed unbiased iterative clustering followed by Harmony-based batch correction (Extended Data Fig. 3a,b), identifying 30 discrete neuronal clusters (Fig. 2a, Extended Data Fig. 3c and Supplementary Data 2). For clarity, these are referred to as 'neuronal clusters' to avoid confusion with the 24 clusters identified in our broad analysis above. Each neuronal cluster was interpreted to represent a unique neuronal cell type or cell state and annotated using gene activity scores for key lineage-defining genes (Fig. 2b and Extended Data Fig. 3d,e). This identified both broad neuronal classes (Extended Data Fig. 3f) and very granular neuronal subdivisions, even discriminating between striatopallidal (neuronal clusters 11 and 12) and striatonigral (neuronal cluster 21) medium spiny neurons, both residing within the striatum but projecting to different brain areas (Fig. 2a and Extended Data Fig. 3g,h). These data identified neuronal cell class-specific peaks, genes and transcription factor activity (Supplementary Fig. 5, Supplementary Data 7 and Supplementary Note 4). While this analysis identified a neuronal cluster corresponding predominantly to substantia nigra dopaminergic neurons (neuronal cluster 7), a key cell type lost in PD, we derived a more refined subset of tyrosine hydroxylase-positive dopaminergic neurons by subclustering only cells from the two substantia nigra samples ($n=403$ dopaminergic neurons; Extended Data Fig. 4a–d).

scATAC–seq pinpoints the cellular targets of GWAS polymorphisms. To understand if any particular cell-type-specific regions of chromatin accessibility were enriched for neurodegenerative disease-associated SNPs, we performed linkage disequilibrium (LD) score regression[41] using a collection of relevant GWAS studies (Supplementary Table 2). Within the peak regions of our broad cell classes, cell-type-specific LD score regression revealed a significant increase in per-SNP heritability for AD in the microglia peak set, reinforcing previous studies[2,42,43] (Fig. 2c and Supplementary Data 8). Similar analyses in PD showed no significant enrichment in SNP heritability in any particular cell type, perhaps because the cellular bases of PD are more heterogeneous than AD (Fig. 2c). Although not a focus of the current study, we note that the data generated can be used to inform the cellular ontogeny of any brain-related GWAS (Fig. 2c). We also confirmed that the heritability of GWAS SNPs from traits not directly related to brain cell types, such as lean body mass and coronary artery disease, was not significantly enriched in any of the tested brain cell types. To ensure that the lack of significance in cell class-specific peaks was not due to obfuscation of neuronal subtypes, we performed the same LD score regression analyses within the peak regions for the neuronal cell classes identified through subclustering (Fig. 2d and Extended Data Fig. 3h). This analysis confirmed our previous findings and showed no significant enrichment for AD or PD SNPs within the peak regions of any neuronal subclasses (Fig. 2d).

Identification of putative enhancer–promoter interactions through chromatin conformation and cell-type-specific coaccessibility. While our scATAC–seq data would enable us to identify the target cell types of functional noncoding SNPs, we sought to additionally identify the target genes of each GWAS locus. To do this, we mapped the enhancer-centric three-dimensional (3D) chromatin architecture in multiple brain regions using Hi-C library preparation followed by chromatin immunoprecipitation (HiChIP)[28] for acetylated histone H3 lysine 27 (H3K27ac), which marks active enhancers and promoters (Extended Data Fig. 5a). In total, we generated 3D interaction maps for 6 of the 7 regions profiled by ATAC–seq (the putamen was excluded given the high overlap with the caudate nucleus), averaging 158 million valid interaction pairs identified per region (Extended Data Fig. 5b,c). We identified 833,975 predicted 3D interactions across all brain regions profiled, of which 331,730 (40%) were reproducible in at least 2 brain regions (Extended Data Fig. 5d and Supplementary Data 9). Of these loops, 67.4% had an ATAC–seq peak present in both anchors, 29.2% had an ATAC–seq peak present in one anchor and 3.4% did not overlap any ATAC–seq peaks identified in either the bulk or scATAC–seq datasets (Extended Data Fig. 5e).

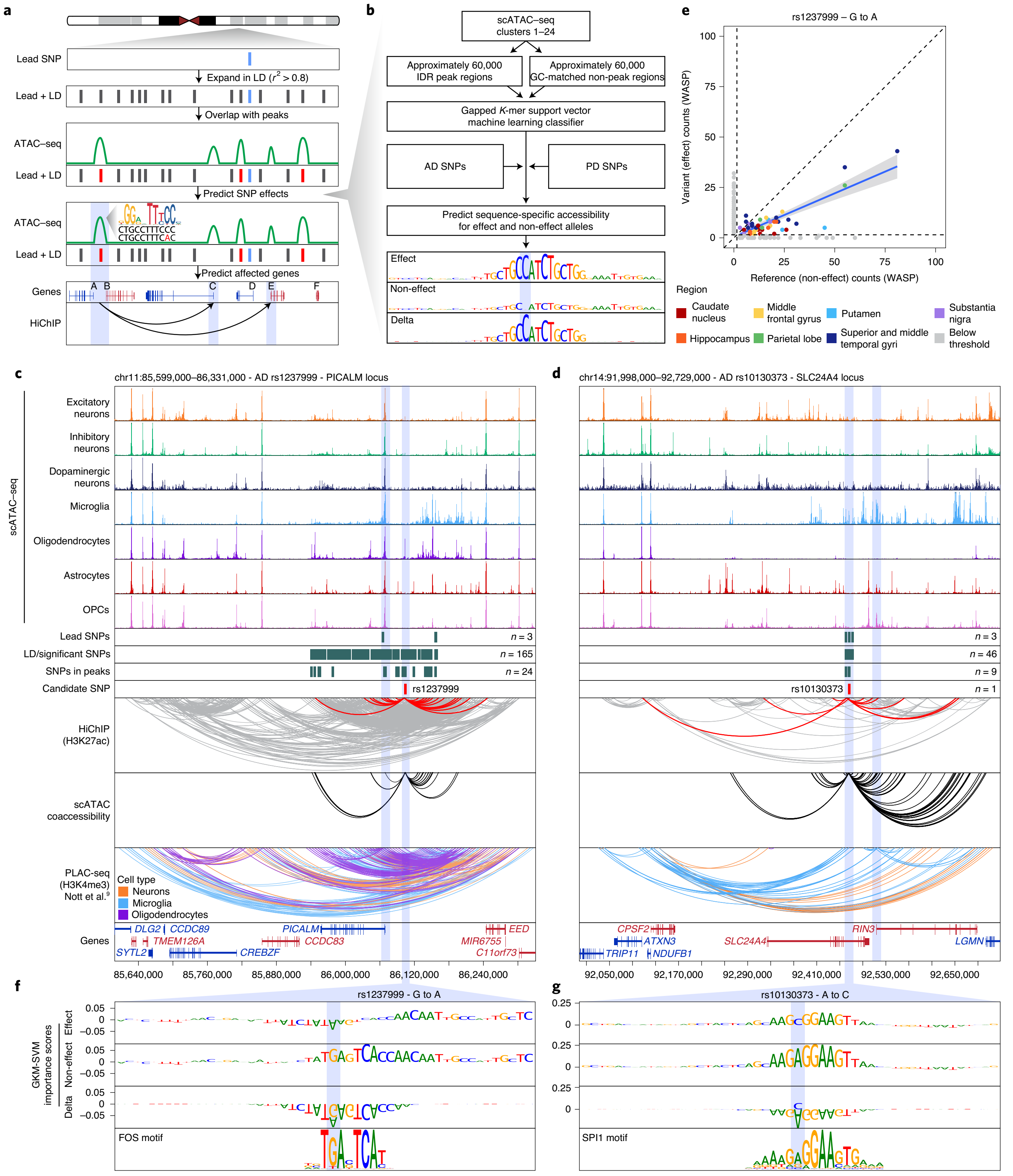
Additionally, correlated variation of chromatin accessibility in peaks across single cells has been shown to predict functional interactions between regulatory elements[31,44]. Using this coaccessibility framework, we predicted regulatory interactions from our scATAC–seq data from the variation across all cells (Extended Data Fig. 5f), identifying 2,822,924 putative pairwise interactions between regions of chromatin accessibility (Supplementary Data 9). This set of interactions showed only moderate overlap (approximately 20%) with our HiChIP data, consistent with the ability of this technique to identify cell-type-specific regulatory interactions, whereas the HiChIP of bulk brain tissue is better suited for the identification of more shared regulatory interactions (Extended Data Fig. 5f,g). Together, these two techniques define a compendium of putative regulatory interactions in the various brain regions studied, thus enabling downstream linkage of GWAS SNPs to putative target genes.

A tiered multi-omic approach to predicting functional noncoding SNPs. To annotate the functional effects of GWAS polymorphisms, we first compiled a comprehensive set of putative disease-relevant SNPs in AD and PD, considering the propensity of nearby SNPs to be coinherited based on LD. We identified (1) any SNPs passing genome-wide significance ($P<5\times10^{-8}$) in recent GWAS[1–3,5–7], (2) any SNPs exhibiting colocalization of GWAS and expression quantitative trait loci signal (FINEMAP/eCAVIAR colocalization posterior probability >0.01) and (3) any SNPs in LD with a SNP in the previous 2 categories based on an LD R^2 value ≥0.8 calculated from phase 1 genotypes of individuals of European ancestry in the 1000 Genomes dataset (Supplementary Table 2 and Methods). In total, this identified 9,707 SNPs including 3,245 unique SNPs across 44 loci associated with AD and 6,496 across 86 loci associated with PD, with a single locus containing 34 SNPs appearing in both diseases.

Fig. 3 | Machine learning predicts functional polymorphisms in AD and PD. a, Schematic of the overall strategy for tiered identification of putative functional SNPs and their corresponding gene targets. **b**, Schematic of the gkm-SVM machine-learning approach used to predict which noncoding SNPs alter transcription factor binding and chromatin accessibility. **c**,**d**, Normalized scATAC–seq-derived pseudobulk tracks, H3K27ac HiChIP loop calls, coaccessibility correlations and publicly available H3K4me3 PLAC-seq loop calls (Nott et al.[9]) in the *PICALM* gene locus (chr11:85,599,000–86,331,000) (**c**) and *SLC24A4* locus (chr14:91,998,000–92,729,000) (**d**). The scATAC–seq tracks represent the aggregate signal of all cells from the given cell type and have been normalized to the total number of reads in TSS regions. For HiChIP, each line represents a FitHiChIP loop call connecting the points at each end. The red lines contain one anchor overlapping the SNP of interest. For coaccessibility, only interactions involving the accessible chromatin region of interest are shown. For PLAC-seq, MAPS loop calls from microglia (blue), neurons (orange) and oligodendrocytes (purple) are shown. **e**, Dot plot showing allelic imbalance at rs1237999. The bulk ATAC–seq counts for the reference/noneffect (G) allele and variant/effect (A) allele are plotted. Each dot represents an individual bulk ATAC–seq sample ($n=140$) colored by brain region. Samples where fewer than three reads were present to support both the reference and variant allele (that is, presumed homozygotes or samples with insufficient sequencing depth) are shown in gray. The blue line represents a linear regression of the non-gray points and the gray box represents the 95% confidence interval of that regression. **f**,**g**, GkmExplain importance scores for each base in the 50-bp region surrounding rs1237999 (**f**) and rs10130373 (**g**) for the effect and noneffect alleles from the gkm-SVM model corresponding to oligodendrocytes (cluster 21) (**f**) and microglia (cluster 24) (**g**). The predicted motif affected by the SNP is shown at the bottom and the SNP of interest is highlighted in blue.

Using this catalog of putative disease-relevant noncoding polymorphisms, we developed a tiered multi-omic approach to predict functional noncoding GWAS polymorphisms by (1) overlapping these SNPs with peaks of chromatin accessibility in our bulk or scATAC–seq data (tier 3), (2) identifying the subset of tier 3 SNPs that may also affect predicted regulatory interactions (tier 2) and (3) predicting which tier 2 SNPs might directly affect transcription factor binding (tier 1) (Fig. 3a and Extended Data Fig. 6a).

To predict these tier 1 SNPs that might directly affect transcription factor binding, we implemented a machine-learning framework to score the allelic effect of a SNP on chromatin accessibility. Using the gapped *k*-mer support vector machine (gkm-SVM)



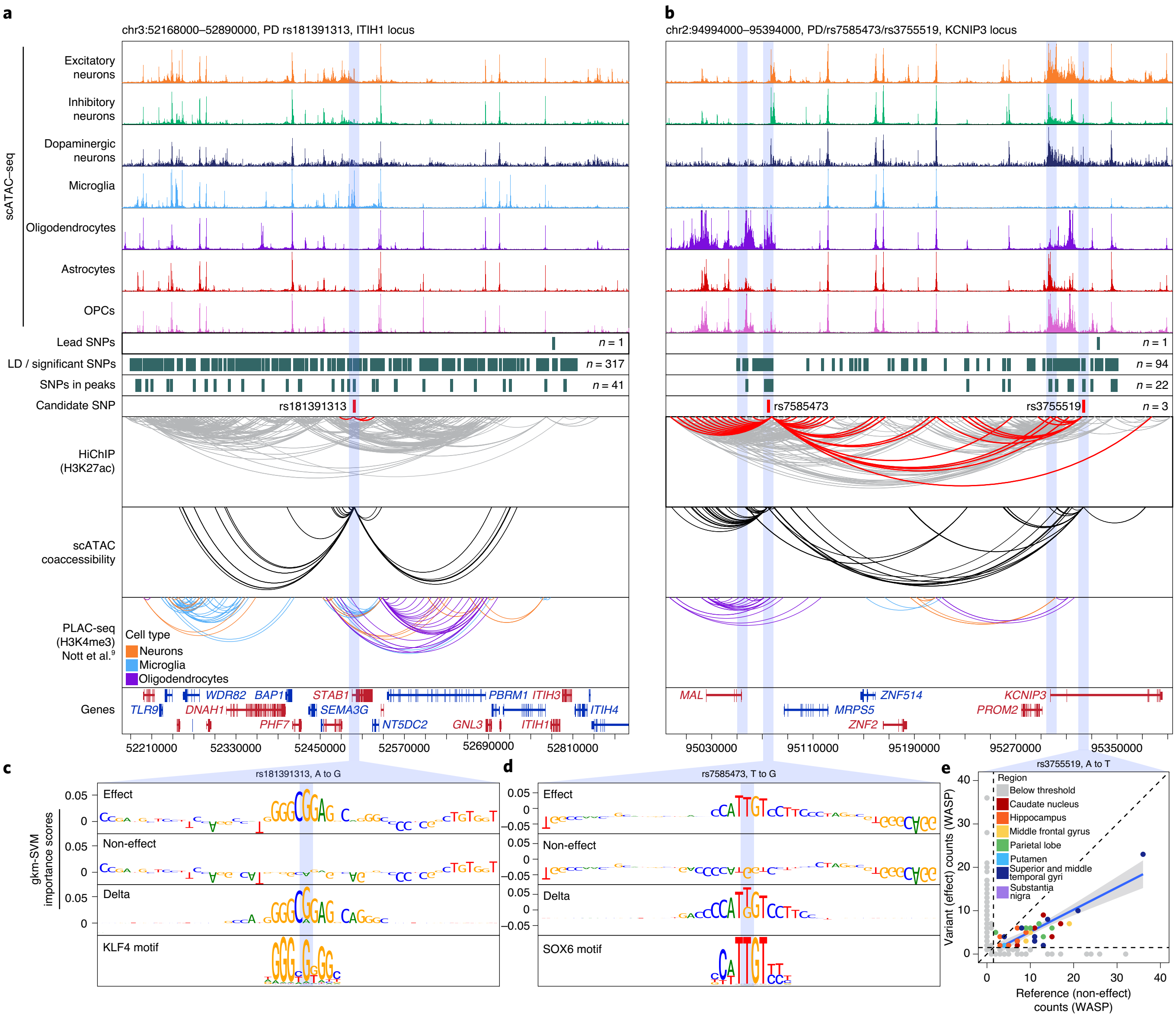


Fig. 4 | Vertical integration of multi-omic data and machine-learning nominate gene targets in AD and PD. a,b, Normalized scATAC-seq-derived pseudobulk tracks, H3K27ac HiChIP loop calls, coaccessibility correlations and publicly available H3K4me3 PLAC-seq loop calls (Nott et al.[9]) in the *ITIH1* gene locus (chr3:52,168,000-52,890,000) (**a**) or the *KCNIP3* locus (chr2:94,994,000-95,394,000) (**b**). The scATAC–seq tracks represent the aggregate signal of all cells from the given cell type and have been normalized to the total number of reads in TSS regions. For HiChIP, each line represents a FitHiChIP loop call connecting the points at each end. The red lines contain one anchor overlapping the SNP of interest. For coaccessibility, only interactions involving the accessible chromatin region of interest are shown. For PLAC-seq, MAPS loop calls from microglia (blue), neurons (orange) and oligodendrocytes (purple) are shown. **c**,**d**, GkmExplain importance scores for each base in the 50-bp region surrounding rs181391313 (**c**) or rs7585473 (**d**) for the effect and noneffect alleles from the gkm-SVM model corresponding to microglia (cluster 24) (**c**) or oligodendrocytes (cluster 21) (**d**). The predicted motif affected by the SNP is shown at the bottom and the SNP of interest is highlighted in blue. **e**, Dot plot showing allelic imbalance at rs3755519. The bulk ATAC–seq counts for the reference/noneffect (A) allele and variant/effect (T) allele are plotted. Each dot represents an individual bulk ATAC–seq sample ($n = 140$) colored by brain region. Samples where fewer than three reads were present to support both the reference and variant allele (that is, presumed homozygotes or samples with insufficient sequencing depth) are shown in gray. The blue line represents a linear regression of the non-gray points and the gray box represents the 95% confidence interval of that regression.

framework[45], we trained predictive regulatory sequence models of chromatin accessibility from each of the 24 broad clusters derived from our scATAC–seq data (Fig. 3b, Supplementary Table 2 and Methods). The gkm-SVM models for all 24 scATAC–seq clusters exhibited high prediction performance on held-out test sequences (Extended Data Fig. 6b,c) and across a tenfold validation scheme (Extended Data Fig. 6d). We used three complementary approaches, GkmExplain[22], in silico mutagenesis[46] and deltaSVM[21] to predict the allelic impact of candidate SNPs on chromatin accessibility in each cluster by providing the sequences corresponding to both alleles of each SNP to the models for each of the 24 clusters. All three approaches showed high concordance of predicted allelic effects across all candidate SNPs (Extended Data Fig. 6e).

As an orthogonal metric for tier 1 SNPs, we performed allelic imbalance analyses with our bulk ATAC–seq data using the robust allele-specific quantitation and quality control (RASQUAL) statistical framework[23] (Extended Data Fig. 6f, Supplementary Data 10 and Methods). Allelic imbalance refers to the differential chromatin accessibility observed between two alleles when one allele is more readily bound by a transcription factor.

Using this tiered approach, we identified genes and molecular processes that could be implicated in AD and PD (Supplementary Fig. 6a–d and Supplementary Note 5). To avoid overinterpretation, we focused our downstream analyses on the subset of GWAS loci that were most likely to involve noncoding regulation based on the absence of any LD SNPs in coding regions (Supplementary Fig. 6e and Supplementary Table 2).

Machine learning predicts putative functional SNPs and identifies the molecular ontogeny of disease associations. This multi-omic approach identified two main categories of new associations within our tier 1 SNPs: (1) established disease-related genes where the precise causative SNP remains unknown; and (2) genes previously not implicated in disease etiology. Many studies have investigated the role of genes such as *PICALM*[47], *SLC24A4* (ref. [48]), *BIN1* (refs. [9,49]) and *MS4A6A*[50] in AD since their implication in the disease by GWAS. However, it is unclear which polymorphisms drive these associations. In the case of *PICALM*, our models predicted a potential functional variant (rs1237999) disrupting a putative FOS/AP1 factor binding site within an oligodendrocyte-specific regulatory element 35 kilobases (kb) upstream of *PICALM* (Fig. 3c,f). Moreover, rs1237999 showed significant allelic imbalance with the variant (effect) allele showing diminished accessibility in bulk ATAC–seq data from heterozygotes across multiple brain regions (Fig. 3e and Supplementary Data 10). Lastly, rs1237999 showed 3D interaction with both *PICALM* and *EED* genes, a polycomb-group family member involved in maintaining a repressive transcriptional state. This expands the potential functional role of this association to a new gene and specifically points to a role for oligodendrocytes, which were not previously implicated in this phenotypic association[47].

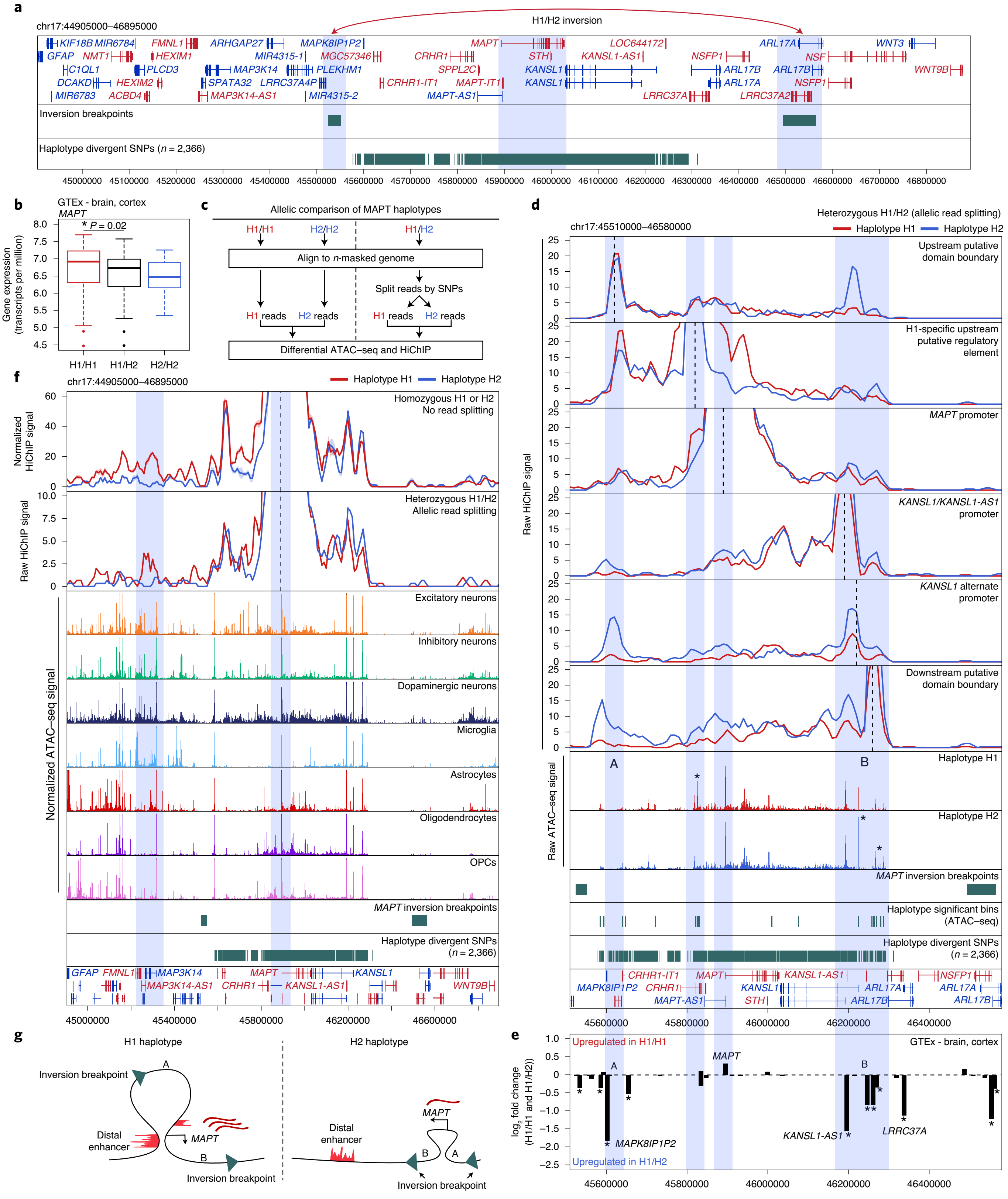
Similarly, the *SLC24A4* locus harbors a small LD block with 46 SNPs that all reside within an intron of *SLC24A4*. Previous work has implicated both *SLC24A4* and the nearby *RIN3* gene in this association but the true mediator is unclear[51,52]. Our multi-omic approach identifies a single SNP, rs10130373, which occurs within a microglia-specific peak, disrupts an SPI1 motif and communicates specifically with the promoter of the *RIN3* gene (Fig. 3d,g). This is consistent with the role of *RIN3* in the early endocytic pathway that is crucial for microglial function and of particular disease relevance in AD[53]. We identified similar examples in the *BIN1* and *MS4A6A* loci (Extended Data Fig. 7 and Supplementary Note 6).

Moreover, the true promise in studying these noncoding polymorphisms is the identification of new genes affected by disease-associated variation. The *ITIH1* GWAS locus occurs within a 600-kb LD block harboring 317 SNPs and no plausible gene association has been made to date. We nominate rs181391313, a SNP occurring within a putative microglia-specific intronic regulatory element of the *STAB1* gene (Fig. 4a). STAB1 is a large transmembrane receptor protein that functions in lymphocyte homing and endocytosis of ligands, such as low-density lipoprotein, two functions consistent with a role for microglia in PD[54]. This SNP is predicted to disrupt a KLF4 binding site, which is consistent with the role of KLF4 in the regulation of microglial gene expression[55] (Fig. 4c). Similarly, the *KCNIP3* GWAS locus resides in a 300-kb LD block harboring 94 SNPs. Our results identify two putative mediators of this phenotypic association with different functional interpretations (Fig. 4b). First, rs7585473 occurs >250 kb upstream of the lead SNP and disrupts an oligodendrocyte-specific SOX6 motif in a peak found to interact with the *MAL* gene, implicated in myelin biogenesis and function (Fig. 4d). Alternatively, we found rs3755519 in a neuronal-specific intronic peak within the *KCNIP3* gene with clear interaction with the *KCNIP3* gene promoter. While this SNP does not show a robust machine-learning prediction, nor reside within a known motif, significant allelic imbalance supports its predicted functional alteration of transcription factor binding (Fig. 4e and Supplementary Data 10). Furthermore, this SNP is associated with *KCNIP3* expression in 3 bulk brain regions from the Genotype-Tissue Expression (GTEx) database (frontal cortex, $P = 4.04 \times 10^{-7}$; hippocampus, $P = 1.45 \times 10^{-7}$; cerebellum, $P = 3.47 \times 10^{-8}$); fine-mapping analysis places rs3755519 within the 95% credible set of causal SNPs in all 3 brain regions. Together, these SNPs provide competing interpretations of this locus, implicating oligodendrocyte- and neuron-specific functions and demonstrating the complexities of interpretation of functional noncoding SNPs. We additionally noted that many SNPs appeared to disrupt binding sites related to CTCF (Extended Data Fig. 8 and Supplementary Note 6).

Epigenomic dissection of the *MAPT* locus explains haplotype-specific changes in local gene expression. One of the strongest PD-associated risk loci is the *MAPT* gene, which encodes tau proteins whose pathological, hyperphosphorylated aggregates form neurofibrillary tangles in AD[56]. However, despite this long-known genetic association, it is unclear how the *MAPT* locus may play a role in PD. The *MAPT* locus is present within a large 1.8-megabase (Mb) LD block and manifests as two distinct haplotypes, H1 and H2, which differ by (1) >2,000 SNPs across the two haplotypes and (2) an approximately 1-Mb inversion that includes the *MAPT* gene[57,58] (Fig. 5a). Previous reports have nominated multiple explanations for how these alterations are associated with

Fig. 5 | Epigenetic deconvolution of the *MAPT* locus explains haplotype-associated transcriptional changes. a, The *MAPT* locus (chr17:44,905,000–46,895,000) showing all genes, the predicted locations of the inversion breakpoints and the 2,366 haplotype-divergent SNPs used for haplotype-specific analyses. **b**, Gene expression of the *MAPT* gene from the GTEx cortex brain samples subdivided based on *MAPT* haplotype ($n = 117$ H1/H1, 78 H1/H2, 10 H2/H2). The lower and upper ends of the box represent the 25th and 75th percentiles and the internal line represents the median. The whiskers represent 1.5 multiplied by the inter-quartile range. Outliers are shown as individual dots. Significance was determined by Wilcoxon rank-sum test. **c**, Schematic for the allelic analysis of the *MAPT* region. **d**, HiChIP (top) and bulk ATAC–seq (middle) sequencing tracks of the region representing the *MAPT* locus inside the predicted inversion breakpoints (chr17:45,510,000–46,580,000; bottom). Each track represents the merge of all available H1 or H2 reads from all heterozygotes. The HiChIP and ATAC–seq tracks represent unnormalized data from heterozygotes where reads were split based on haplotype. HiChIP is shown as a virtual 4C plot where the anchor is indicated by a dotted line and the signal represents paired-end tag counts overlapping a 10-kb bin. Regions showing significant haplotype bias in ATAC–seq are marked with an asterisk (Wilcoxon rank-sum test). **e**, GTEx cortex gene expression of genes in the *MAPT* locus comparing H1 homozygotes ($n = 117$) to H1/H2 ($n = 78$). Regions A and B are shown as in Fig. 5d. *$P < 0.05$ by Wilcoxon rank-sum test after multiple hypothesis correction. **f**, HiChIP (top) and cell-type-specific scATAC–seq (middle) sequencing tracks of the region representing the *MAPT* locus outside of the predicted inversion breakpoints (bottom). HiChIP tracks for bulk homozygote H1 or H2 samples (normalized based on reads-in-loops) are shown at the top while haplotype-specific tracks from heterozygotes (unnormalized) are shown below. In each HiChIP plot, the anchor represents the *MAPT* promoter. scATAC–seq tracks represent the aggregate signal of all cells from the given cell type and have been normalized to the total number of reads in TSS regions. **g**, Schematic illustrating the predicted haplotype-specific change in long-distance interaction between the *MAPT* promoter and the predicted distal regulatory element identified in Fig. 5d. Regions marked A and B represent the same regions marked in Fig. 5d,e.

PD, including increased *MAPT* expression in the H1 haplotype[59,60] (Fig. 5b), different ratios of splice isoforms[61–63] and the use of alternative promoters[64]. We created a haplotype-specific map of chromatin accessibility and 3D chromatin interactions at the *MAPT* locus (Fig. 5c). Using data from heterozygote H1/H2 individuals, we split reads into H1 and H2 haplotypes based on the presence of 1 of the 2,366 haplotype-divergent SNPs (Supplementary Table 2 and Methods). We tiled the region into non-overlapping 500-base pair (bp) bins (to avoid biases in peak calling) and performed a Wilcoxon rank-sum test to identify regions differentially accessible both between H1/H1 and H2/H2 homozygotes and between split reads from H1/H2 heterozygotes (Extended Data Fig. 9a,b). This



identified 28 differentially accessible bins including an H1-specific putative regulatory element located 68 kb upstream of the *MAPT* promoter and the promoter of the *KANSL1* gene located 330 kb downstream of *MAPT* (Fig. 5d (asterisks) and Extended Data Fig. 9c). Using our HiChIP data, we performed haplotype-specific virtual 4C to determine if any changes in chromatin accessibility were accompanied by changes in 3D chromatin interaction frequency. We identified H2-specific 3D interactions between a putative domain boundary upstream of *MAPT* (labeled 'A') and the region surrounding the *KANSL1* promoter (labeled 'B') spanning a distance of >600 kb inside the inversion breakpoints (Fig. 5d). Additionally, the H1-specific putative regulatory element upstream of *MAPT* showed increased interaction with a second putative regulatory element intronic to *MAPT* and with the *MAPT* promoter (Fig. 5d).

To better understand how these epigenetic changes impact haplotype-specific gene expression, we used RNA sequencing data from the GTEx database. In addition to the previously mentioned haplotype-specific differences in *MAPT* expression (Fig. 5b), we also identified significant changes in gene expression near the largest changes in chromatin accessibility and 3D interaction ('A' and 'B'; Fig. 5e and Extended Data Fig. 9d,e). These increases in gene expression could play a functional role in *MAPT* haplotype-mediated pathological changes or, more likely, be a nonfunctional by-product of the genomic inversion.

These analyses illuminate how the genomic region inside the *MAPT* inversion breakpoints differs between the H1 and H2 haplotypes; alternatively, the inversion could alter *MAPT* gene expression by changing the relative orientation of the *MAPT* gene to enhancers and promoters outside of the breakpoints. In support of this, we identified a long-distance putative regsulatory element located 650 kb upstream of the *MAPT* gene that showed elevated interaction with the *MAPT* promoter specifically in the H1 haplotype (Fig. 5f). Indeed, we found multiple neuron-specific putative regulatory elements in this upstream region, consistent with the known neuron-specific expression of *MAPT* (Extended Data Fig. 9f), and an increase in overall 3D interaction between this upstream region and the region surrounding *MAPT* inside of the inversion breakpoints (Extended Data Fig. 9g). Additional studies are needed to demonstrate the functional effects of these predicted regulatory interactions (Fig. 5g).

Discussion

In this study, we provide a high-resolution epigenetic characterization of the role of inherited noncoding variation in AD and PD. Our integrative multi-omic framework and machine-learning classifier predicted dozens of functional SNPs, nominating gene and cellular targets for each noncoding GWAS locus. These predictions both inform well-studied disease-relevant genes, such as *BIN1* in AD, and suggest new gene-disease associations, such as *STAB1* in PD. This expands our understanding of inherited variation in AD and PD and provides a roadmap for epigenomic dissection of noncoding variation in neurodegenerative and other complex genetic diseases.

Together, this multi-omic resource captures the regional and cellular gene regulatory machinery that governs phenotypic expression of noncoding variation, thus allowing us to identify most polymorphisms that could putatively affect gene expression through overlap with peaks of chromatin accessibility (tier 3). To further refine these putative functional variants, we identified the subset of polymorphisms that could be mapped to gene targets through 3D chromatin interactions or coaccessibility networks (tier 2). Finally, we employed a machine-learning approach to predict the subset of polymorphisms likely to perturb transcription factor binding and validated these predictions with measurements of allelic imbalance (tier 1). In total we implicate approximately five times as many genes in the phenotypic association of AD and PD and nominate functional noncoding variants for dozens of previously orphaned GWAS loci. Additionally, through our integrative analysis, we provide a comprehensive epigenetic characterization of the *MAPT* gene locus (discussed in detail in Supplementary Note 7). The functional predictions made through our machine-learning classifier and integrative analytical approach greatly expand our understanding of noncoding contributions to AD and PD. More broadly, this work represents a systematic approach to understanding inherited variation in disease and provides an avenue towards the nomination of new therapeutic targets that previously remained obscured by the complexity of the regulatory machinery of the noncoding genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00721-x.

Received: 27 December 2019; Accepted: 18 September 2020;
Published online: 26 October 2020

References

1. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
2. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
3. Lambert, J.-C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
4. Beecham, G. W. et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS Genet.* **10**, e1004606 (2014).
5. Pankratz, N. et al. Meta-analysis of Parkinson's disease: identification of a novel locus, *RIT2*. *Ann. Neurol.* **71**, 370–384 (2012).
6. Chang, D. et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).
7. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
8. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
9. Nott, A. et al. Brain cell type-specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
10. Li, M. et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, eaat7615 (2018).
11. Amiri, A. et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science* **362**, eaat6720 (2018).
12. Trevino, A. E. et al. Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367**, eaay1645 (2020).
13. Nowakowski, T. J. et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
14. Song, M. et al. Mapping *cis*-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019).
15. Rajarajan, P. et al. Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science* **362**, eaat4311 (2018).
16. Fullard, J. F. et al. An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).
17. Fullard, J. F. et al. Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci. *Hum. Mol. Genet.* **26**, 1942–1951 (2017).
18. Bryois, J. et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nat. Commun.* **9**, 3121 (2018).
19. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
20. Sey, N. Y. A. et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nat. Neurosci.* **23**, 583–593 (2020).
21. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
22. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: fast and accurate interpretation of nonlinear gapped *k*-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).

23. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
24. Amlie-Wolf, A. et al. INFERNO: inferring the molecular mechanisms of noncoding genetic variants. *Nucleic Acids Res.* **46**, 8740–8753 (2018).
25. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
26. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
27. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
28. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
29. Mumbach, M. R. et al. Enhancer connectome in primary human cells reveals target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
30. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
31. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
32. Corces, M. R. et al. Lineage-specific and single cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
33. Corces, M. R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
34. McKeown, M. R. et al. Superenhancer analysis defines novel epigenomic subtypes of non-APL AML, including an RARα dependency targetable by SY-1425, a potent and selective RARα agonist. *Cancer Discov.* **7**, 1136–1153 (2017).
35. Stolt, C. C. et al. The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes Dev.* **17**, 1677–1689 (2003).
36. Kuhlbrodt, K., Herbarth, B., Sock, E., Hermans-Borgmeyer, I. & Wegner, M. Sox10, a novel transcriptional modulator in glial cells. *J. Neurosci.* **18**, 237–250 (1998).
37. Kondo, T. & Raff, M. Basic helix-loop-helix proteins and the timing of oligodendrocyte differentiation. *Development* **127**, 2989–2998 (2000).
38. Nakatani, H. et al. Ascl1/Mash1 promotes brain oligodendrogenesis during myelination and remyelination. *J. Neurosci.* **33**, 9752–9768 (2013).
39. Smith, A. M. et al. The transcription factor PU.1 is critical for viability and function of human brain microglia. *Glia* **61**, 929–942 (2013).
40. Schlingensiepen, K. H. et al. The role of Jun transcription factor expression and phosphorylation in neuronal differentiation, neuronal cell death, and plastic adaptations in vivo. *Cell. Mol. Neurobiol.* **14**, 487–505 (1994).
41. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
42. Hemonnot, A. L., Hua, J., Ulmann, L. & Hirbec, H. Microglia in Alzheimer disease: well-known targets and new opportunities. *Front. Aging Neurosci.* **11**, 233 (2019).
43. Efthymiou, A. G. & Goate, A. M. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener*. **12**, 43 (2017).
44. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
45. Ghandi, M. et al. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).
46. Bromberg, Y. & Rost, B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. *Bioinformatics* **24**, i207–i212 (2008).
47. Xu, W., Tan, L. & Yu, J.-T. The role of PICALM in Alzheimer's disease. *Mol. Neurobiol.* **52**, 399–413 (2015).
48. Stage, E. et al. The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimers Dement. (Amst)* **5**, 53–66 (2016).
49. Andrew, R. J. et al. Reduction of the expression of the late-onset Alzheimer's disease (AD) risk-factor *BIN1* does not affect amyloid pathology in an AD mouse model. *J. Biol. Chem.* **294**, 4477–4487 (2019).
50. Ma, J., Yu, J.-T. & Tan, L. MS4A cluster in Alzheimer's disease. *Mol. Neurobiol.* **51**, 1240–1248 (2015).
51. Rouka, E. et al. Differential recognition preferences of the three Src homology 3 (SH3) domains from the adaptor CD2-associated protein (CD2AP) and direct association with Ras and Rab interactor 3 (RIN3). *J. Biol. Chem.* **290**, 25275–25292 (2015).
52. Larsson, M. et al. GWAS findings for human iris patterns: associations with variants in genes that influence normal neuronal pattern development. *Am. J. Hum. Genet.* **89**, 334–343 (2011).
53. Kajiho, H. et al. RIN3: a novel Rab5 GEF interacting with amphiphysin II involved in the early endocytic pathway. *J. Cell Sci.* **116**, 4159–4168 (2003).
54. Lecours, C. et al. Microglial implication in Parkinson's disease: loss of beneficial physiological roles or gain of inflammatory functions? *Front. Cell. Neurosci.* **12**, 282 (2018).
55. Kaushik, D. K., Gupta, M., Das, S. & Basu, A. Krüppel-like factor 4, a novel transcription factor regulates microglial activation and subsequent neuroinflammation. *J. Neuroinflammation* **7**, 68 (2010).
56. Schellenberg, G. D. & Montine, T. J. The genetics and neuropathology of Alzheimer's disease. *Acta Neuropathol.* **124**, 305–323 (2012).
57. Stefansson, H. et al. A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
58. Zody, M. C. et al. Evolutionary toggling of the *MAPT* 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
59. Valenca, G. T. et al. The role of *MAPT* haplotype H2 and isoform 1N/4R in Parkinsonism of older adults. *PLoS ONE* **11**, e0157452 (2016).
60. Allen, M. et al. Association of *MAPT* haplotypes with Alzheimer's disease risk and *MAPT* brain gene expression levels. *Alzheimers Res. Ther.* **6**, 39 (2014).
61. Pascale, E. et al. Genetic architecture of *MAPT* gene region in Parkinson disease subtypes. *Front. Cell. Neurosci.* **10**, 96 (2016).
62. Beevers, J. E. et al. *MAPT* genetic variation and neuronal maturity alter isoform expression affecting axonal transport in iPSC-derived dopamine neurons. *Stem Cell Reports* **9**, 587–599 (2017).
63. Lai, M. C. et al. Haplotype-specific *MAPT* exon 3 expression regulated by common intronic polymorphisms associated with Parkinsonian disorders. *Mol. Neurodegener.* **12**, 79 (2017).
64. Huin, V. et al. Alternative promoter usage generates novel shorter *MAPT* mRNA transcripts in Alzheimer's disease and progressive supranuclear palsy brains. *Sci. Rep.* **7**, 12589 (2017).

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

Methods

Publicly available data used in this study. All QTL analysis was performed using GTEx v.8. Additionally, we downloaded full-genome summary statistics of GWAS associations for three AD cohorts[1–3] and two PD cohorts[6,65]; however, these cohorts are not all mutually exclusive. The PD full GWAS summary statistics from Chang et al.[6] were obtained through a research agreement with 23andMe. These summary statistics included those generated by 23andMe ($n = 6{,}476$ individuals with PD and 302,042 controls who were disease-free) but not summary statistics from individuals incorporated into the meta-analysis from the original publication. All GWAS data used in this study (except the data protected through our research agreement with 23andMe) have been compiled for ease of reproducibility and are available at https://doi.org/10.1101/2020.01.06.896159 and https://zenodo.org/record/3817811. Additionally, we obtained MAPS-based loop calls directly from published proximity ligation-assisted ChIP–seq (PLAC-seq) data from microglia, neurons and oligodendrocytes[9].

Genome annotations. All data are aligned and annotated to the hg38 reference genome.

Sequencing. Bulk ATAC–seq and HiChIP were sequenced using an Illumina HiSeq 4000 System with paired-end 75-bp reads. scATAC–seq was sequenced using an Illumina NovaSeq 6000 System with an S4 flow cell with paired-end 99-bp reads.

Sample acquisition and patient consent. Primary brain samples were acquired postmortem with institutional review board-approved informed consent from Stanford University, the University of Washington or Banner Health. Human donor sample sizes were chosen to provide sufficient confidence to validate methodological conclusions. Human brain samples were collected with an average postmortem interval of 3.9 h (range 2.0–6.9 h). These brain regions include distinct isocortical regions (superior and middle temporal gyri, Brodmann areas 21 and 22), parietal lobe (Brodmann area 39) and middle frontal gyrus (Brodmann area 9), striatum at the level of the anterior commissure (caudate nucleus and putamen), hippocampus at the level of the lateral geniculate nucleus and the substantia nigra at the level of the red nucleus. Macrodissected brain regions were flash-frozen in liquid nitrogen. Some samples were embedded in optimal cutting temperature compound. All samples were stored at −80 °C until use. Due to the limiting nature of these primary samples, this unique biological material is not available upon request.

Isolation of nuclei from frozen tissue chunks and bulk ATAC–seq data generation. Nuclei were isolated from frozen tissue as described previously[19,33]. This protocol, including the transposition reaction, is available on protocols.io (https://doi.org/10.17504/protocols.io.6t8herw). Briefly, frozen tissue fragments were Dounce-homogenized to create a suspension of nuclei. Nuclei were purified using an iodixanol gradient and washed in resuspension buffer (RSB). Nuclei were counted and, for each replicate, 50,000 nuclei were aliquoted into a separate tube containing RSB with 0.1% Tween-20. Nuclei were pelleted and transposed as described in the protocol linked above according to the Omni-ATAC transposition conditions[19]. Transposed fragments were purified and amplified as described previously[26] with slight modification. Briefly, transposed fragments were preamplified for three cycles. The concentration of preamplified fragments was determined by quantitative PCR (qPCR) and this concentration was used to estimate the total number of cycles required to obtain 160 fmol of fragments. A second PCR was performed to amplify the preamplified fragments for the desired number of cycles. Final libraries were again purified. Before sequencing, libraries were pooled and run on a 6% polyacrylamide gel electrophoresis gel and excess primers and primer dimers below 125 bp were removed. Libraries were sequenced on an Illumina HiSeq 4000 System as described above. After isolation and bulk ATAC–seq, the remaining nuclei were cryopreserved in BAMBANKER (Wako Chemicals) and stored at −80 °C for use in other assays such as scATAC–seq and HiChIP.

Statistics. All statistical tests performed are included in the figure legends or Methods where relevant.

ATAC–seq data processing. The ENCODE Data Coordination Center (DCC) ATAC–seq pipeline (https://doi.org/10.5281/zenodo.211733) (v.1.1.7) was used to process the bulk ATAC–seq samples, starting from FASTQ files. The pipeline was executed with irreproducible discovery rate (IDR) enabled and the IDR threshold was set to 0.05. The GRCh38 reference genome assembly was used, keeping only the primary chromosomes chr1–chr22, chrX, chrY, chrM. The pipeline was executed with ATAQC enabled, using GENCODE v.29 transcription start site (TSS) annotations. Biological replicates were analyzed individually, with the two technical replicates for each biological replicate provided as inputs to the 'atac.bams' argument of the pipeline. Other arguments to the pipeline were kept at their defaults.

ATAC–seq peak calling. Pipeline peak calls underwent several levels of filtering to identify credible peak sets. The IDR optimal peak set from the DCC pipeline for each biological replicate was determined. Although the IDR peaks for individual biological replicates were corrected for multiple testing, the high number of biological samples in the dataset served as another source of multiple testing error. To address this source of error, tagAlign files for all biological replicates for a given brain region/condition were concatenated. The DCC pipeline (v.1.1.7) was subsequently executed on the merged tagAlign files as single-replicate inputs. The pipeline generated pseudoreplicates from the input tagAlign files for each brain region/condition. Optimal IDR peaks were called from the pseudoreplicates. This set of IDR peaks was filtered to keep peaks supported by 30% or more of the IDR peaks from the pipeline runs on individual biological replicates.

Sample-by-peak count matrices were then generated from the resulting set of filtered peaks. Filtered peaks from the pooled tagAlign files were concatenated and truncated to within 200 bp of the summit (100-bp flank kept upstream and downstream of the peak summit). These 200-bp regions were merged with the BEDTools[66] (v.2.26.0) merge command to avoid merging peaks with low levels of overlap. BEDTools coverage was used to compute the number of tagAlign reads that overlapped each peak region in the pseudoreplicates in the merged tagAlign dataset. This analysis yielded a total of $n = 186{,}559$ peaks combined across the brain regions.

Motif enrichment. Motif enrichment was performed using the hypergeometric test as described previously[33,67].

Feature binarization. Identification of 'unique' peaks from the ATAC–seq data was performed as described previously[33]. Briefly, for each of the cell classes (termed 'groups'), we created three pseudobulk replicates that were used to create a count matrix of insertion counts within each peak of the scATAC–seq peak set. This count matrix was then log-normalized using 'edgeR::cpm(mat,log = TRUE,prior.count = 3)'. We then calculated the intragroup mean and intragroup s.d. across every peak in the scATAC–seq peak set. Then, for each peak, we ranked the groups by their intragroup mean. Then, we iterated from the second lowest group asking whether the mean of that group was greater than the maximum intragroup mean plus the intragroup s.d. of the next-lowest sample. This iterative process proceeds until a group was identified that met this criterion. This point was defined as the breakpoint and all groups with a higher intragroup mean were classified as positive for this peak and given a value of '1'. All groups below the breakpoint were given a value of '0'. If a peak did not have a breakpoint, it was discarded. This peak 'binarization' procedure classified all '1s' as being higher than every individual '0'. This also captured the peaks that were unique to multiple groups. We kept all combinations that were unique to three or fewer groups. To facilitate multiple hypothesis testing, we computed a contrast matrix for all observed combinations and ran limma's (v.3.38.3) eBayes test on the log-normalized counts matrix. We then extracted all of the false discovery rate (FDR)-adjusted *P* values from differential testing keeping those peaks that were below an FDR of 0.001. This resulted in the classification of 221,062 peaks.

Sequencing tracks. Sequencing tracks were created using the WashU Epigenome Browser. All sequencing tracks of a given locus have the same *y* axis. All tracks show data that have been normalized by 'reads-in-peaks' (for ATAC–seq) or 'reads-in-loops' for HiChIP to account for differences in signal-to-background ratios across multiple samples, unless otherwise stated. For all sequencing tracks, genes that are on the plus strand (that is, 5′ to 3′ in the left to right direction) are shown in red and genes that are on the minus strand (that is, 5′ to 3′ in the right to left direction) are shown in blue to enable identification of the TSS.

LD score regression. We apply stratified LD score regression, a method for partitioning heritability from GWAS summary statistics, to sets of cell-type-specific ATAC–seq peaks to identify disease-relevant cell types for AD and PD along with other brain-related GWAS traits. Using our scATAC–seq data, peak coordinates were first converted from hg38 to hg19 for analysis with GWAS data. We followed the LD score regression tutorial (https://github.com/bulik/ldsc/wiki) as used previously[41] for single-cell-specific analysis[68]. We used brain-related GWAS summary statistics, such as AD[1], PD[6], schizophrenia[69], anorexia nervosa[70], attention deficit hyperactivity disorder[71], anxiety[72], neuroticism[73] and epilepsy[74] (Supplementary Table 2 and https://zenodo.org/record/3817811). To serve as controls, we also used summary statistics for GWAS of traits not obviously linked to brain tissues such as lean body mass[75], bone mineral density[76] and coronary artery disease[77]. In particular, we looked at the regression coefficient *P* value, indicative of the contribution of this annotation to trait heritability, conditional on the baseline model described previously[41].

Allele counts from ATAC–seq data. The WASP mapping pipeline (https://github.com/bmvdgeijn/WASP/tree/master/mapping) was used to reduce biases in mapping and filtering duplicate reads. Reads were mapped using bowtie2 (v.2.3.2) to the UCSC hg38 reference genome. Variants were called on the resulting BAM files using bcftools mpileup v.1.9 to produce VCF files. These VCF files and the WASP-corrected BAM files were used as input for the GATK ASEReadCounter (v.4.0.1.2) tool to obtain allele counts and their mapping quality. These allele counts were used to visualize significant allelic imbalance as determined by RASQUAL (https://github.com/natsuhiko/rasqual). For plotting, samples that lacked at least three read counts for both the reference and alternate alleles were inferred to be either homozygous or too low coverage to presume heterozygosity. However, these allele counts were only used for display purposes and did not contribute to any determination of significance for allelic imbalance.

Allelic imbalance from ATAC–seq data using RASQUAL. We intersected the coordinates of all LD-expanded candidate AD and PD GWAS and colocalization SNPs with peaks from our ATAC–seq data to obtain the candidate SNPs that we tested for allele-specific effects on chromatin accessibility. We used the createASVCF.sh script from the RASQUAL[23] GitHub repository (https://github.com/natsuhiko/rasqual) to obtain the allele-specific counts at each candidate SNP for all samples. We used the fitAseNullMulti function from the QuASAR[78] (v.0.1) GitHub repository to calculate for each donor the posterior probability of the three possible genotypes at all of the candidate SNP positions using all available brain region samples from that donor and assigned the genotype at each position to be the one with the highest posterior probability. Next, using these allele-specific counts and genotypes and the allele frequencies from the 1000 Genomes Project[79] for each candidate SNP, we created a VCF file for each brain region, which included the allele-specific counts and genotypes from only the samples that originated from those respective regions. Similarly, we created region-specific count matrices, which contained columns of ATAC–seq read counts for each feature only from the samples that originated from the respective regions. We also ran the makeOffset.R script from the RASQUAL repository with a list of guanine-cytosine (GC) contents, corresponding to the GC content of each feature in the counts matrix, as an argument to generate the sample-specific offset terms file for each brain region. Since RASQUAL is run on each feature from the counts matrix independently of other features, we further split the region-specific input VCF files, counts matrices and offset files by chromosome and used the text2bin.R script from the RASQUAL repository to convert the region and chromosome-specific input counts matrices and offset files into the binary format required by RASQUAL.

Finally, we ran RASQUAL using the input VCF file, counts matrix and offset file from each of the 22 chromosomes (chromosomes 1–22; chromosome X and chromosome Y did not have any candidate SNPs) from each of the brain regions and tested each candidate SNP present in each feature in the counts matrix. To test for genome-wide significance of each putative chromatin accessibility QTL, we ran RASQUAL with the–random-permutation option along with the same inputs ten times to generate a background set of null q values. For each brain region, we used the empirical distribution of null q values to identify those SNPs that have a q value lower than the 10% FDR threshold as significant chromatin accessibility QTLs, as recommended by the authors (https://github.com/natsuhiko/rasqual/issues/21).

Selection of candidate SNPs for ATAC–seq overlap analysis, HiChIP interaction tests and gkm-SVM model-based allelic effect scores. Our goal was to identify SNPs with a causal effect on any of the selected GWAS traits. To minimize the chances of excluding causal GWAS SNPs, we selected the set of all variants achieving a genome-wide significant $P<5\times10^{-8}$ for any GWAS trait. We then added in any lead SNPs from the colocalization analysis that achieved a colocalization posterior probability score >0.01, even those that did not pass the genome-wide significance value of $P<5\times10^{-8}$. We also included all trait-associated SNPs curated from two other PD studies[6,7]. In these studies, full summary statistics were not publicly available for the entire genome because the meta-analysis was applied only to the subset of SNPs reaching genome-wide significance in a previous PD GWAS. We then computed the full set of SNPs that had an LD $R^2\geq0.8$ with at least 1 of the SNPs in the set selected above. These LD calculations were performed on the phase 1 genotypes of individuals of European ancestry in the 1000 Genomes dataset, provided in full at https://zenodo.org/record/3404275#.Xlw62XVKhhE. Pairwise LD values of all variants in the above subset were calculated via PLINK v.1.90. These pairwise LD values were used to identify 1000 Genomes SNPs with $R^2\geq0.8$ with the SNPs in our dataset. Together, these LD buddies plus the original set of trait-relevant SNPs comprised the set of SNPs tested in our subsequent functional analyses.

Testing GWAS loci for overlap with ATAC–seq peaks. We tested all SNPs in the above set for overlap with ATAC–seq peaks from two different annotation formats. The first annotation consisted of bulk ATAC–seq peaks identified in one of seven brain regions. The second annotation consisted of cluster-specific peaks from scATAC–seq data. For each variant selected for functional analysis, we determined all cellular contexts where an ATAC–seq peak contained this variant, as well as the nearest peak if no peak contained the variant.

scATAC–seq library generation. Cryopreserved nuclei were thawed on ice and 65,000 nuclei were transferred to a tube containing 1 ml of RSB-T (10 mM of Tris-HCl pH 7.5, 10 mM of NaCl, 3 mM of $MgCl_2$ and 0.1% Tween). Nuclei were pelleted at 500 RCF for 5 min at 4 °C in a fixed-angle rotor. The supernatant was fully removed using two pipetting steps (p1000 to remove down to the last 100 µl, then p200 to remove all remaining supernatant). This pellet was then gently resuspended in 12 µl of 1× nuclei buffer (10x Genomics). To transpose, 5 µl of this nuclei suspension (containing 27,000 nuclei) was transferred to a tube containing 10 µl of transposition mix (10x Genomics). This reaction mixture was incubated at 37 °C for 1 h to transpose. The remainder of library generation was completed as described in the 10x Genomics Single Cell ATAC Regent Kits User Guide (v.1 Chemistry).

scATAC–seq latent semantic indexing (LSI) clustering and visualization. scATAC–seq clustering analysis was performed using an alpha version of the ArchR software[80]. To cluster our scATAC–seq data (for both broad clustering and neuronal subclustering), we first identified a robust set of peak regions followed by iterative LSI clustering[27,30]. Briefly, we created 1-kb windows tiled across the genome and determined whether each cell was accessible within each window (binary). Next, we identified the top 50,000 accessible windows across all samples (accounting for GC bias) and performed an LSI dimensionality reduction (term frequency-inverse document frequency (TF-IDF) transformation followed by singular value decomposition (SVD)) on these windows followed by Harmony batch correction[81]. We then performed Seurat[82] clustering (FindClusters v.2.3) on the harmonized LSI dimensions at resolutions of 0.8, 0.4 and 0.2, keeping the clustering for which the minimum cluster size was greater than 100 cells (0.2 if this condition was not met). For each cluster, we called peaks on the Tn5-corrected insertions (each end of the Tn5-corrected fragments) using the MACS2 callpeak command with the parameters '--shift --75 --extsize 150 --nomodel --call-summits --nolambda --keep-dup all -q 0.05'. The peak summits were then extended by 250 bp on either side to a final width of 501 bp, filtered by the ENCODE hg38 blacklist (https://www.encodeproject.org/annotations/ENCSR636HFF/) and filtered to remove peaks that extended beyond the ends of chromosomes. We then created a non-overlapping set of extended summits across all of these peaks as described previously[27,30].

We then counted the accessibility for each cell in these peak regions to create an accessibility matrix. We then adopted the iterative LSI clustering approach[27,30] to unbiasedly identify clusters that were due to biological versus technical variation. Briefly, we computed the TF-IDF transformation as described by Cusanovich et al.[83]. To do this, we divided each index by the colSums of the matrix to compute the cell 'term frequency'. Next, we multiplied these values by log(1 + ncol(matrix)/rowSums(matrix)), which represents the 'inverse document frequency'. This yields a TF-IDF matrix that can be used as input to irlba's (v.2.3.3) SVD implementation in R (v.3.6.1). We then used Harmony to batch-correct the LSI dimensions in R. Using the first 25 reduced dimensions as input into a Seurat object, crude clusters were identified using Seurat's (v.2.3) shared nearest neighbor (SNN) graph clustering FindClusters function with a resolution of 0.2. We then calculated the cluster sums from the binarized accessibility matrix and then log-normalized them using edgeR's 'cpm(matrix, log = TRUE, prior.count = 3)' in R. Next, we identified the top 25,000 varying peaks across all clusters using the 'rowVars' function in R. This was done on the cluster log-normalized matrix rather than the sparse binary matrix because: (1) it reduced biases due to cluster cell sizes; and (2) it attenuated the mean variability relationship by converting to log space with a scaled prior count. The 25,000 variable peaks were then used to subset the sparse binarized accessibility matrix and recompute the TF-IDF transform. We used SVD on the TF-IDF matrix to generate a lower dimensional representation of the data by retaining the first 25 dimensions. We then used Harmony to batch-correct the LSI dimensions in R. We then used these reduced dimensions as input into a Seurat object and crude clusters were identified using Seurat's SNN graph clustering FindClusters function with a resolution of 0.6. This process was repeated a third time with a resolution of 1.0. Then, these same reduced dimensions were used as input to Seurat's 'RunUMAP' function with default parameters and plotted in ggplot2 (v.3.2.1) using R.

scATAC–seq gene activity scores. Gene activity scores are based on the observation that chromatin accessibility within the gene body, at the promoter and at distal regulatory elements is correlated with gene expression[30,31,80,84]. Gene scores were calculated using ArchR v.0.9.4 (ref. [80]) with default parameters. Briefly, ArchR infers gene activity scores using a distance-weighted accessibility model that aggregates the accessibility signal inside the gene body and in the local genomic region. The resulting gene activity scores were additionally imputed using MAGIC[85] (v.2.0.3) to reduce noise due to scATAC–seq data sparsity.

Identification of clusters and cell types from scATAC–seq data. Different clusters and cell types were manually identified using promoter accessibility and gene activity scores for various lineage-defining genes. Microglia (cluster 24) were identified based on accessibility near the *IBA1*, *CD14*, *CD11C*, *PTGS1* and *PTGS2* genes. Astrocytes (clusters 13–17) were identified based on accessibility near the *GFAP* and *FGFR3* genes. Excitatory neurons (clusters 1, 3 and 4 were identified based on accessibility near the *SLC17A6* and *SLC17A7* genes. Inhibitory neurons (clusters 2, 11 and 12) were identified based on accessibility near the *GAD2* and *SLC32A1* genes. Medium spiny neurons (most of cluster 2) were identified based on accessibility near the *PPP1R1B* gene. Oligodendrocytes (clusters 19–23) were identified based on accessibility near the *MAG* and *SOX10* genes. OPCs (clusters 8–10) were identified based on accessibility near the *PDGFRA* gene. All neuronal subsets were identified primarily as neurons based on accessibility near the *NEFL*, *RBFOX3*, *VGF* and *GRIN1* genes and then subdivided based on the region of origin and accessibility near the other genes mentioned above.

scATAC–seq peak calling. For scATAC–seq peak calling from clusters or manually defined cell types, all single cells belonging to the given group were pooled together. These pooled fragment files were converted to the paired-end tagAlign format and processed with v.1.4.2 of the ENCODE DCC ATAC–seq pipeline. The conversion to tagAlign was performed as follows. For fragments on the positive strand, the read start coordinate was the fragment start coordinate, zero-indexed. The read end coordinate was the fragment start coordinate plus the read length (99 bp). For fragments on the negative strand, the read start coordinate was the fragment end coordinate, zero-indexed. The read start coordinate was the fragment end coordinate minus the read length (99 bp). Then, these tagAlign files were used as input to the DCC ATAC–seq pipeline. IDR optimal peak sets with an IDR threshold of 0.05 were determined for each cluster by the pipeline, using pseudobulk replicate tagAligns for the cluster. Other pipeline parameters were the same as for bulk ATAC–seq data (see above).

scATAC–seq pseudobulk replicate generation and differential accessibility comparisons. For differential comparisons of clusters or cell types, including Pearson correlation determination, non-overlapping pseudobulk replicates were generated from groups of cells. For each cell grouping (that is, a cluster or a cell type), a minimum of 300 cells was required to make at least two non-overlapping pseudobulk replicates of 150 cells each. A maximum of 3 pseudobulk replicates was made per group if the total number of cells per group was >450 cells. Cells were randomly deposited into one of the pseudobulk replicates and all available cells were used. In this way, the non-overlapping pseudobulk replicates were agnostic to which donor the cell came from but aware of individual cells (that is, all reads from a given cell were deposited into the same pseudobulk replicate). These pseudobulk replicates were then used for differential comparisons using DESeq2 (v.1.24.0) (ref. [86]).

Identification of neuronal cell class-specific peaks, transcription factor motifs and genes. ArchR was used to call peaks (using addReproduciblePeakSet) and identify cell class-specific peaks and genes (using getMarkerFeatures). The cell class-specific peaks were tested for motif enrichment (using peakAnnoEnrichment).

Transcription factor footprinting. Transcription factor footprinting was performed as described previously[33].

HiChIP library generation. HiChIP library generation was performed as described previously[28]. One million cryopreserved nuclei were used per experiment. Enzyme MboI (New England Biolabs) was used for restriction digestion. Sonication was performed on a Covaris E220 instrument using the following settings: duty cycle = 5; peak incident power = 140; cycles per burst = 200; time 4 min. All HiChIP was performed using H3K27ac as the target (catalog no. ab4729; Abcam).

HiChIP data analysis. HiChIP paired-end sequencing data were processed using HiC-Pro[87] v.2.11.0 with a minimum mapping quality of 10. FitHiChIP[88] was used to identify 'peak-to-all' interactions using peaks called from the one-dimensional HiChIP data. A lower distance threshold of 20 kb and an upper distance threshold of 2 Mb were used. Bias correction was performed using coverage-specific bias.

HiChIP linkage of SNPs to genes. To link SNPs to genes, we identified FitHiChIP loops that contained a SNP in one anchor and a TSS in the other anchor. This was performed for all LD-expanded SNPs to identify the full complement of genes that could be putatively implicated in AD and PD.

gkm-SVM machine-learning classifier training and testing. See the Supplementary Methods.

Identification of *MAPT* haplotypes. The *MAPT* haplotype block is part of one of the largest LD blocks in the human genome. To identify SNPs that belonged exclusively to either the H1 or H2 haplotype, we used minor allele frequencies from dbSNP v.151. SNPs were required to be within the coordinates of the *MAPT* inversion breakpoints (hg38 chr17:45,551,578–46,494,237) and have a minor allele frequency between 8.4 and 9%. While there are undoubtedly haplotype-specific SNPs outside this frequency range, we chose this range to be as conservative as possible and pick SNPs that showed minimal haplotype switching. Each SNP was verified to track with the predicted haplotype using LDlink[89] (v.3.6). This resulted in 2,366 SNPs that could be confidently called as haplotype divergent.

***MAPT* locus differential expression analysis.** A 900-kb block of variants in strong LD at the *MAPT* locus hampered the resolution of colocalization methods for identifying causal variants and/or genes at this locus. To probe this locus more dee ply, we assembled a list of 2,366 variants uniquely found in either the H1 or the H2 haplotype of the *MAPT* locus (described above). For each of the 838 individuals genotyped in GTEx, we counted the number of variants in support of either haplotype. We designated individuals as homozygous if they possessed <1% of variants favoring the opposite haplotype and heterozygous if 45–55% of variants supported either haplotype. This determined the individual's haplotype in all but six cases, which were excluded from the remainder of the *MAPT* analysis. In total, we identified 539 individuals with the H1/H1 haplotype, 260 with H2/H1 and 33 with H2/H2. Our a priori gene of interest was *MAPT*, whose expression had previously been demonstrated to be higher in H1 than H2 haplotypes. At a nominal cutoff of $P < 0.05$, we confirmed this expected direction of differential *MAPT* expression (higher in H1 haplotypes) in multiple tissues, with the strongest contrasts in 'Brain—Cortex'.

We then extended our analysis to include all genes expressed in any of the brain tissues from GTEx. We compared the $\log_2$ fold change of gene expression (transcripts per million) between H1/H1 and H1/H2 individuals, given that these subgroups had the largest sample size. A change was considered statistically significant if a Wilcoxon rank-sum test between the two groups produced a $P < 0.05$/(total no. of genes)/(total no. of tissues). We also performed pairwise Wilcoxon rank-sum test comparisons for each gene in each brain tissue between all three pairings of haplotypes.

***MAPT* haplotype-specific ATAC–seq and HiChIP analysis.** For both ATAC–seq and HiChIP, reads from heterozygote donors were remapped to an *n*-masked genome (using bowtie2 or HiC-Pro, respectively) where all dbSNP positions were masked to '*n*'. After alignment, SNPsplit[90] (v.0.3.2) was used to divide reads mapping to either the H1 or H2 haplotypes based on the presence of one of the 2,366 haplotype-divergent SNPs identified above. In this way, reads mapping to regions that lack a haplotype-divergent SNP could not be assigned in an allelic fashion to either the H1 or H2 haplotypes and were ignored. For track-based visualizations of haplotype-specific data, all available data from a given haplotype were merged agnostically to the brain region the data were derived from. For visualization of ATAC–seq and HiChIP data from H1/H2 heterozygotes, no normalization was performed because each sample was internally controlled for allelic depth. To identify regions with haplotype-specific chromatin accessibility in the *MAPT* locus, the entire locus was tiled into non-overlapping 500-bp bins and the number of Tn5 transposase insertions were counted for each haplotype in each bin for each sample. A Wilcoxon signed-rank test was used to determine if the difference between H1 and H2 for each bin was significant after multiple hypothesis correction (FDR < 0.01).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data generated in this work are available through GEO accession no. GSE147672. To facilitate wide access to our data, we created a WashU Epigenome browser session (session ID: drS3o1n4kJ) for our scATAC–seq data in the following track formats: (1) broad cell types (Corces_scATAC_BroadCellTypes);(2) broad clusters (Corces_scATAC_BroadClusters); (3) neuron subclusters (Corces_scATAC_NeuronSubClusters); and (4) neuron subclustered cell types/LDSC groups (Corces_scATAC_NeuronSubCellTypes). These tracks are accessible via the following link: http://epigenomegateway.wustl.edu/legacy/?genome=hg38&session=drS3o1n4kJ.

Code availability

All custom code used in this work is available at the following GitHub repository: https://github.com/kundajelab/alzheimers_parkinsons.

References

65. Pankratz, N. et al. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.* **124**, 593–605 (2009).
66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
67. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
68. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
69. Li, Z. et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).
70. Duncan, L. et al. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *Am. J. Psychiatry* **174**, 850–858 (2017).
71. Demontis, D. et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63–75 (2019).
72. Otowa, T. et al. Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry* **21**, 1391–1399 (2016).
73. Okbay, A. et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).

74. Anney, R. J. L. et al. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **13**, 893–903 (2014).
75. Zillikens, M. C. et al. Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun.* **8**, 80 (2017).
76. Kemp, J. P. et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).
77. Howson, J. M. M. et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* **49**, 1113–1119 (2017).
78. Harvey, C. T. et al. QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics* **31**, 1235–1242 (2015).
79. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
80. Granja, J. M. et al. ArchR: an integrative and scalable software package for single-cell chromatin accessibility analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.28.066498 (2020).
81. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
82. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
83. Cusanovich, D. A. et al. The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
84. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
85. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).
86. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
87. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
88. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.* **10**, 4221 (2019).
89. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
90. Krueger, F. & Andrews, S. R. SNPsplit: allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res.* **5**, 1479 (2016).

Acknowledgements

This work was supported by National Institutes of Health (NIH) grant nos. NS062684, AG057707, AG053959, AG047366 (to T.J.M.), HG007735 (to H.Y.C.), HG009431 (to S.B.M./A.K.), AG066490 (to S.B.M.) and AG059918 (to M.R.C.). Additional support for patient sample collection was provided by NIH grant nos. AG005136 and AG019610. Sequencing data for this project were generated on an Illumina HiSeq 4000 System supported in part by NIH grant no. S10OD018220. Additional resources at the Stanford Center for Genomics and Personalized Medicine Sequencing Center were supported by NIH grant no. S10OD025212. M.R.C. is supported by the American Society of Hematology Scholar Award. A.S. is supported by the Stanford BioX Bowes fellowship. M.J.G. and T.E. are supported by a National Library of Medicine training grant no. 5T15LM007033-36. M.J.G. is additionally supported by a Stanford Graduate Fellowship. H.Y.C. is an Investigator of the Howard Hughes Medical Institute.

Author contributions

M.R.C., H.Y.C. and T.J.M. conceived of and designed the project. M.R.C. and T.J.M. compiled the figures and wrote the manuscript with help and input from all authors. A.S. and M.R.C. performed the bulk ATAC–seq data processing and analysis. M.R.C. performed all HiChIP data analysis with help from M.R.M. and J.M.G. J.M.G., M.R.C. and A.S. performed all scATAC–seq data processing and analysis with supervision from W.J.G., A.K., S.B.M. and H.Y.C. M.J.G. performed the GWAS locus curation, colocalization analysis and GTEx analysis. M.J.G., L.F. and B.L. performed all LD score regression analysis with supervision from S.B.M. S.K. and A.S. performed the machine-learning analysis with supervision from A.K. S.K. and T.E. performed the allelic imbalance analyses with supervision from A.K. and S.B.M. B.H.L., S.S. and M.R.C. performed all ATAC–seq, scATAC–seq and HiChIP data generation with help from S.T.B. and M.R.M. K.S.M. curated the frozen tissue specimens used in this work.

Competing interests

H.Y.C. is a cofounder of Accent Therapeutics, Boundless Bio, and an advisor to 10x Genomics, Arsenal Bio and Spring Discovery. S.B.M. is on the scientific advisory board of MyOme. A.K. is a consultant for Biogen. A.S. is a consultant for MyoKardia. W.J.G. is a consultant for Guardant Health, 10x Genomics and Protillion Biosciences.

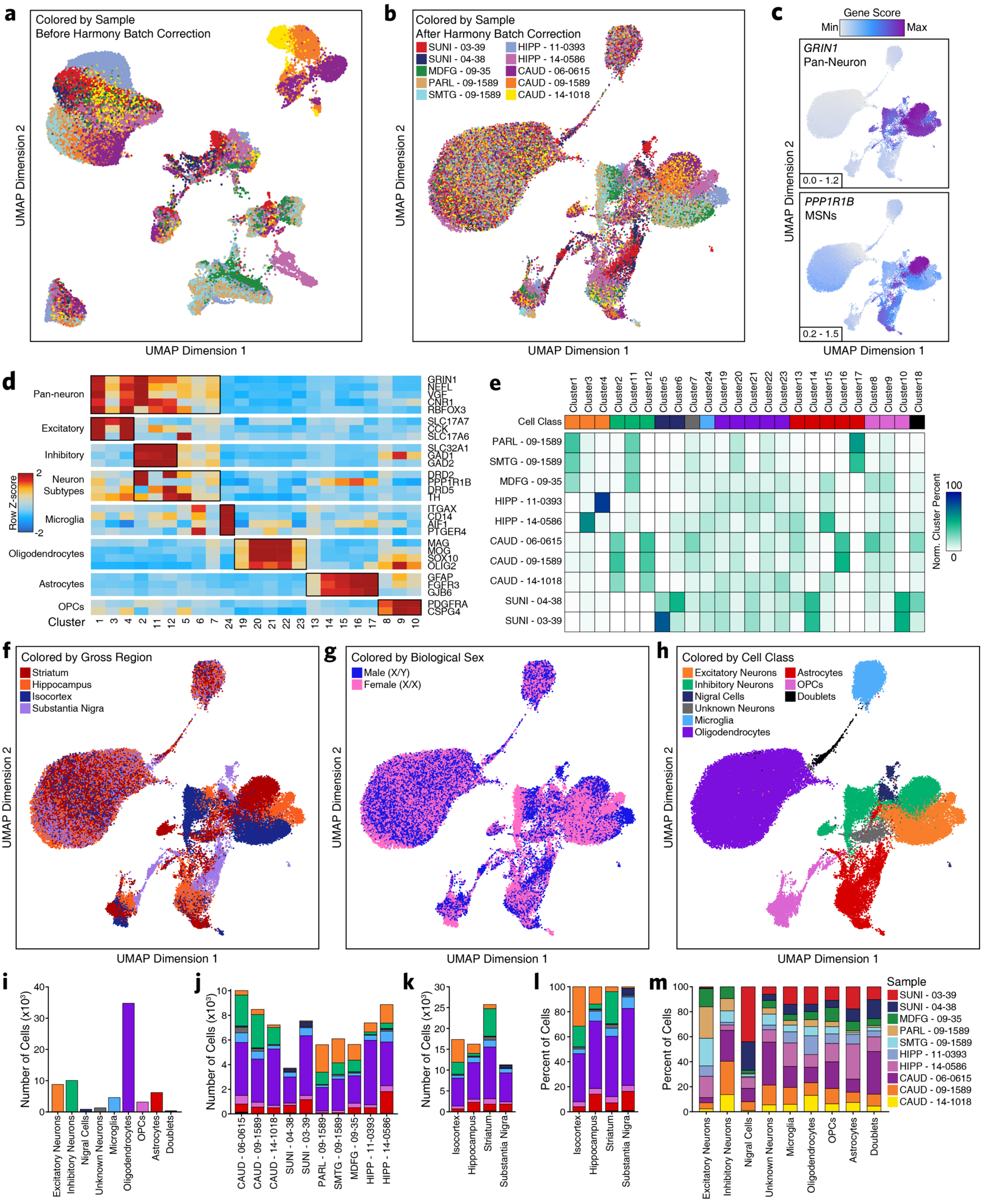
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00721-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00721-x.

Correspondence and requests for materials should be addressed to H.Y.C. or T.J.M.

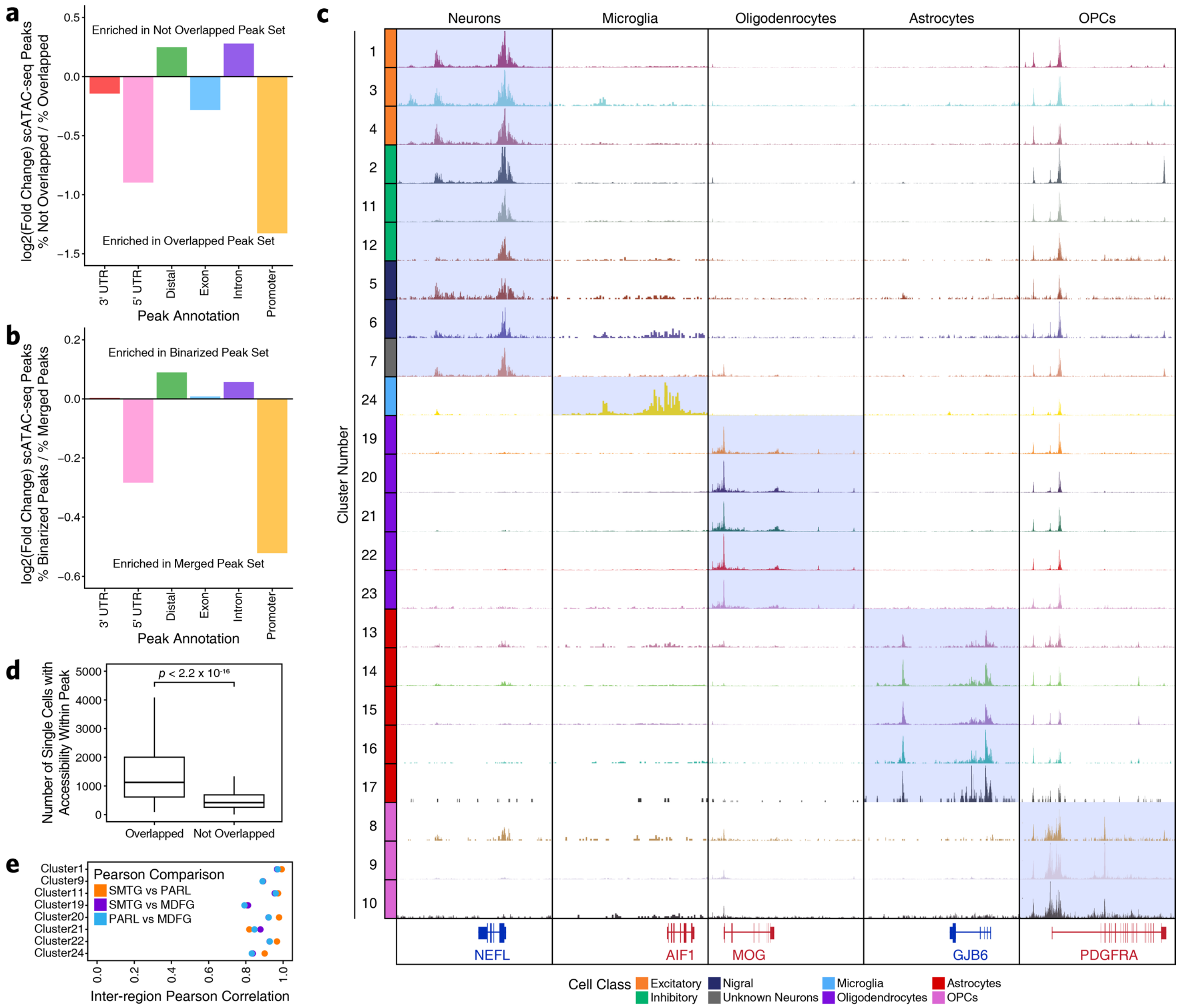
Reprints and permissions information is available at www.nature.com/reprints.



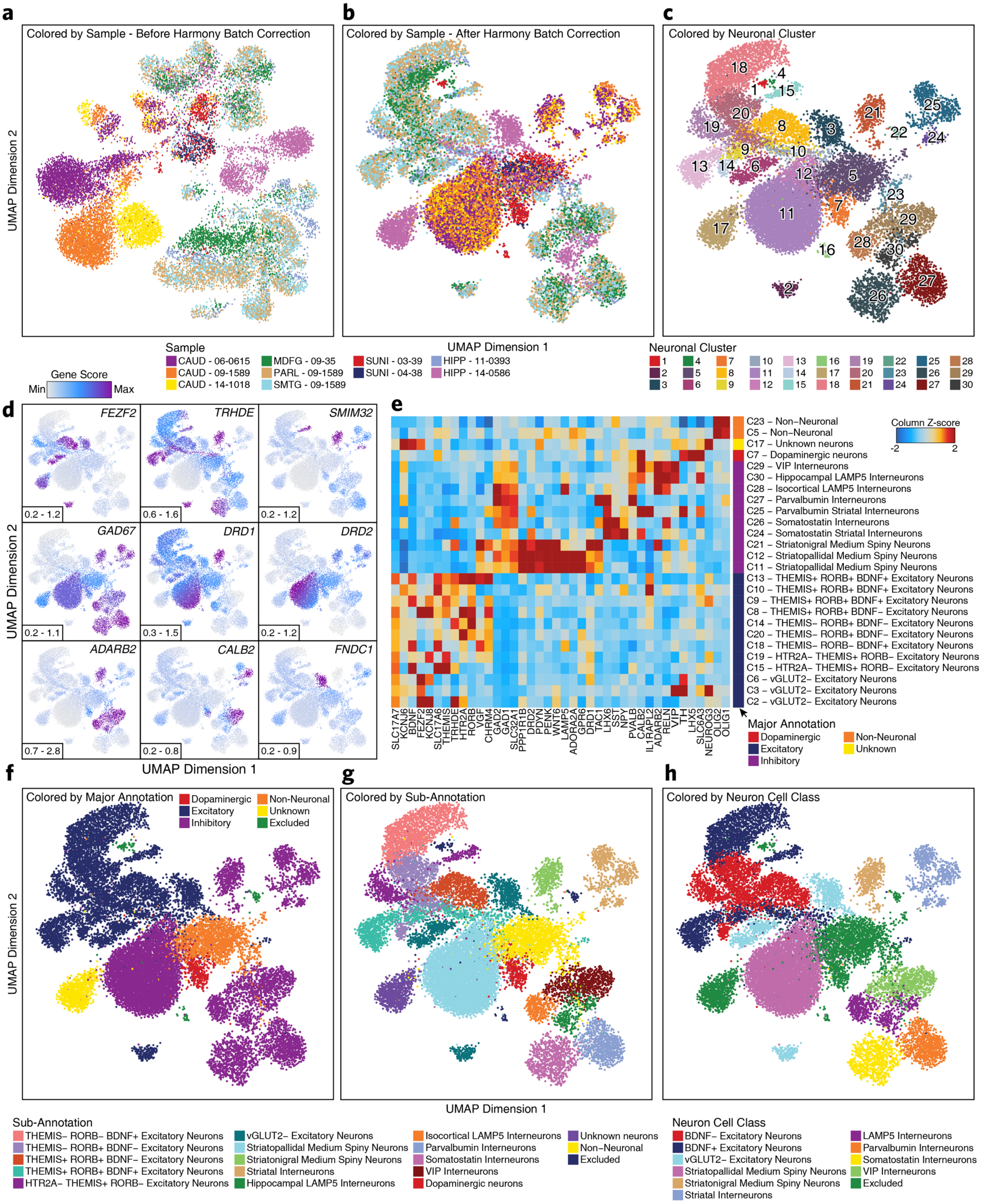
Extended Data Figure 1

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Region-centric scATAC-seq identifies cellular and regional heterogeneity in chromatin accessibility in adult brain. a, **b**, UMAP dimensionality reduction (**a**) prior to and (**b**) after batch correction with Harmony of scATAC-seq data from 10 different samples. Each dot represents a single cell (N = 70,631). Dots are colored by the sample of origin. Color labels are shown in Extended Data Fig. 1b. **c**, The same UMAP dimensionality reduction shown in Extended Data Fig. 1b but each cell is colored by its gene activity score for the annotated lineage-defining gene. Gene activity scores were imputed using MAGIC. Grey represents the minimum gene activity score while purple represents the maximum gene activity score for the given gene. The minimum and maximum scores are shown in the bottom left of each panel. The gene of interest and the cell type that it identified are shown in the upper left of each panel. MSNs – medium spiny neurons. **d**, Heatmap of cell type-specific markers used to define the cell type corresponding to each cluster. Color represents the row-wise Z-score of chromatin accessibility in the vicinity of each gene for each cluster. **e**, Cluster residence heatmap showing the percent of each cluster that is composed of cells from each sample. Cell numbers were normalized across samples prior to calculating cluster residence percentages to account for differences in total pass filter cells per sample. **f**–**h**, UMAP dimensionality reduction as shown in Extended Data Fig. 1b but colored by (**f**) the gross brain region from which each cell was obtained, (**g**) the biological sex of the donor for each cell, or (**h**) the predicted cell class for each cell. **i**–**k**, Bar plot showing the number of cells identified in our scATAC-seq data from (**i**) each of the annotated cell classes, (**j**) each of the annotated donors/samples, or (**k**) each of the gross brain regions subdivided based on cell class. Color represents the predicted cell class as shown in the legend of Extended Data Fig. 1h. **l**, **m**, Bar plot showing the percentage of cells in our scATAC-seq data from (**l**) each of the gross brain regions subdivided based on cell class or (**m**) each of the annotated cell classes subdivided based on donor/sample of origin. Color represents (**l**) the predicted cell class as shown in the Extended Data Fig. 1h or (**m**) the biological sample from which the cells were obtained.

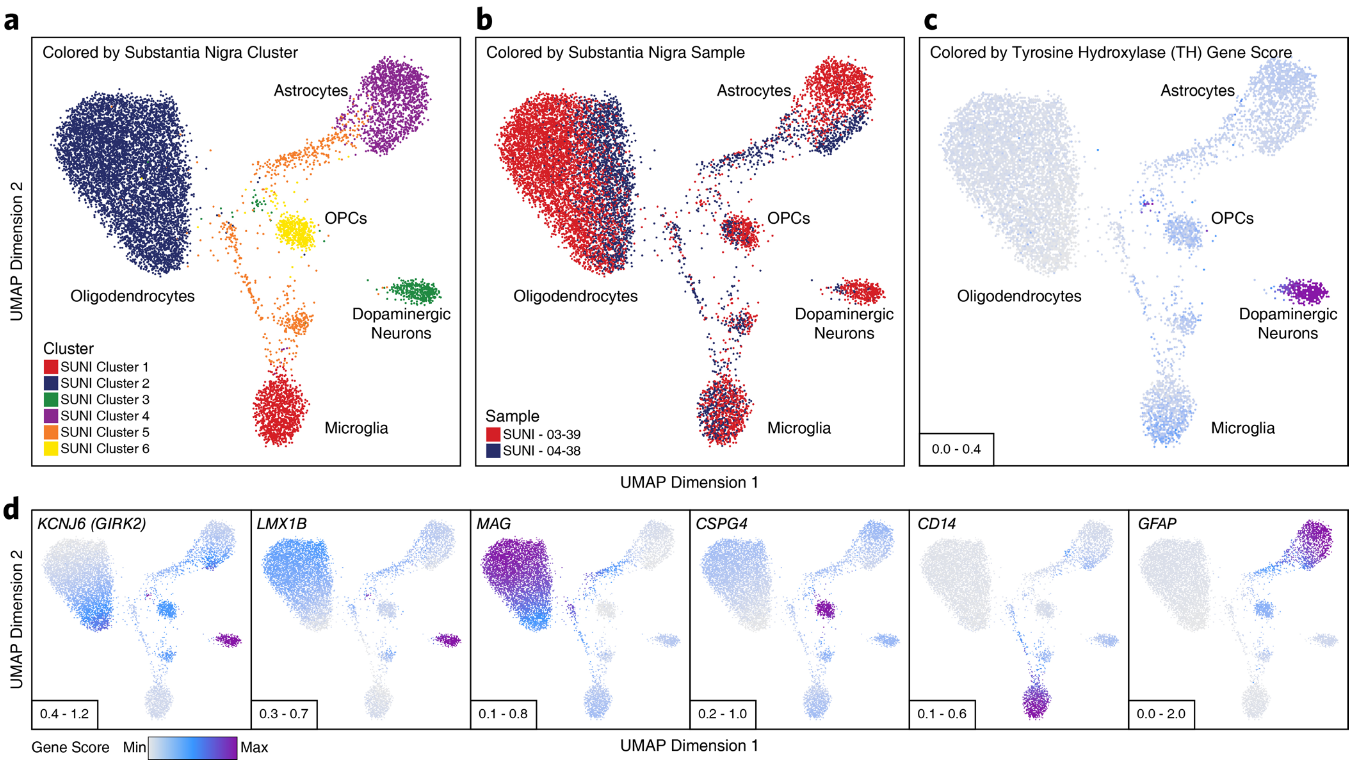


Extended Data Fig. 2 | Cellular heterogeneity in brain tissue necessitates single-cell approaches to capture biological complexity. a, **b**, Bar plot of the log2(Fold Change) in the percent of peaks mapping to various genomic annotations comparing peaks from (**a**) the scATAC-seq peak set that are not overlapped by a peak from the bulk ATAC-seq peak set to peaks that are overlapped by a peak from the bulk ATAC-seq peak set or (**b**) the scATAC-seq peak set that were identified as cell type-unique through feature binarization to all peaks from the scATAC-seq peak set. **c**, Sequencing tracks of lineage-defining factors shown across all 24 scATAC-seq clusters (except Cluster 18 – putative doublets). From left to right, *NEFL* (neurons; chr8:24933431-24966791), *AIF1* (aka *IBA1*, microglia; chr6:31607841-31617906), *MOG* (oligodendrocytes; chr6:29652183-29699713), *GJB6* (astrocytes; chr13:20200243-20239571), and *PDGFRA* (OPCs; chr4:54209541-54303643). **d**, Box and whiskers plots showing the distribution of the number of single cells from our scATAC-seq data showing accessibility within (left) each peak from the set of peaks from the scATAC-seq peak set that overlap a peak from the bulk ATAC-seq peak set (N = 120,941 peaks) and (right) each peak from the set of peaks from the scATAC-seq peak set that do not overlap a peak from the bulk ATAC-seq peak set (N = 238,081 peaks). The lower and upper ends of the box represent the 25th and 75th percentiles and the internal line represents the median. The whiskers represent 1.5 multiplied by the inter-quartile range. P-value determined by Kolmogorov–Smirnov test. **e**, Dot plot showing the inter-region Pearson correlation of pseudo-bulk replicates comprised of all cells from either SMTG, PARL, or MDFG within each of the clusters shown. The clusters shown were selected based on biological relevance (that is clusters annotated as "substantia nigra astrocytes" should not be compared across isocortical regions) and on cluster size (that is clusters with small numbers of isocortical cells would not provide robust comparisons).

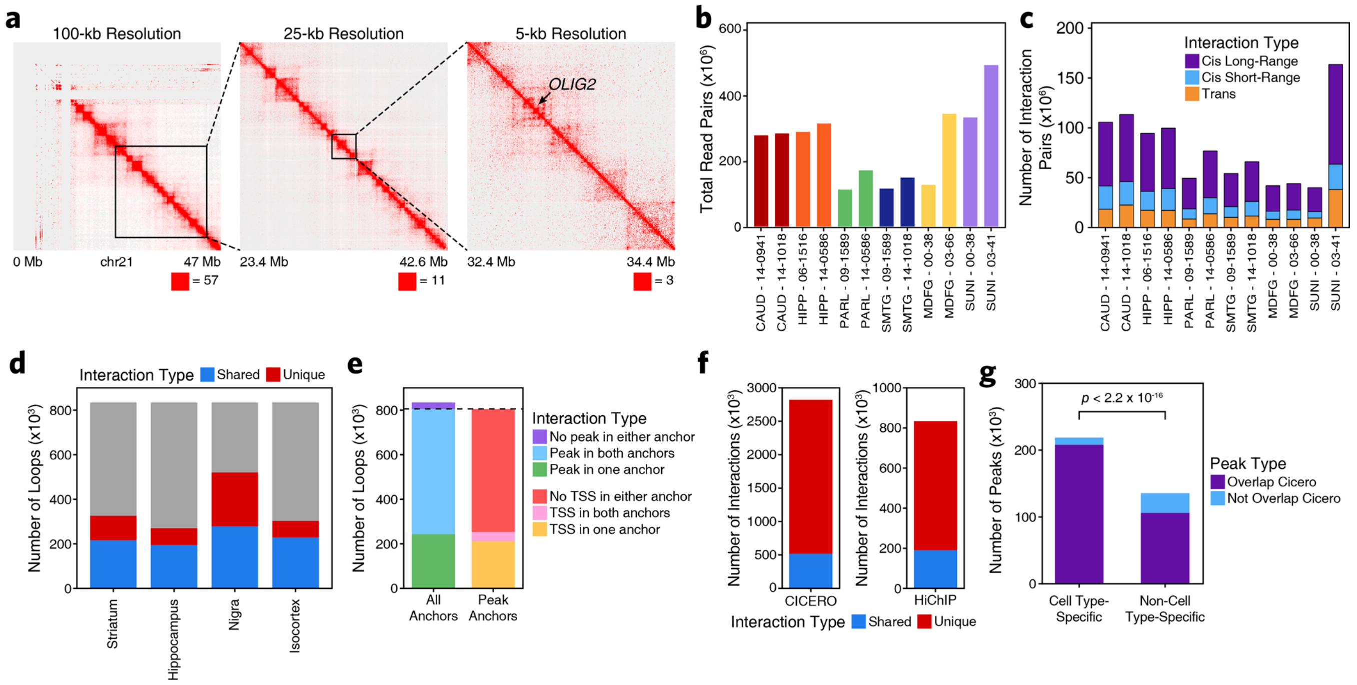


Extended Data Fig. 3 | See next page for caption.

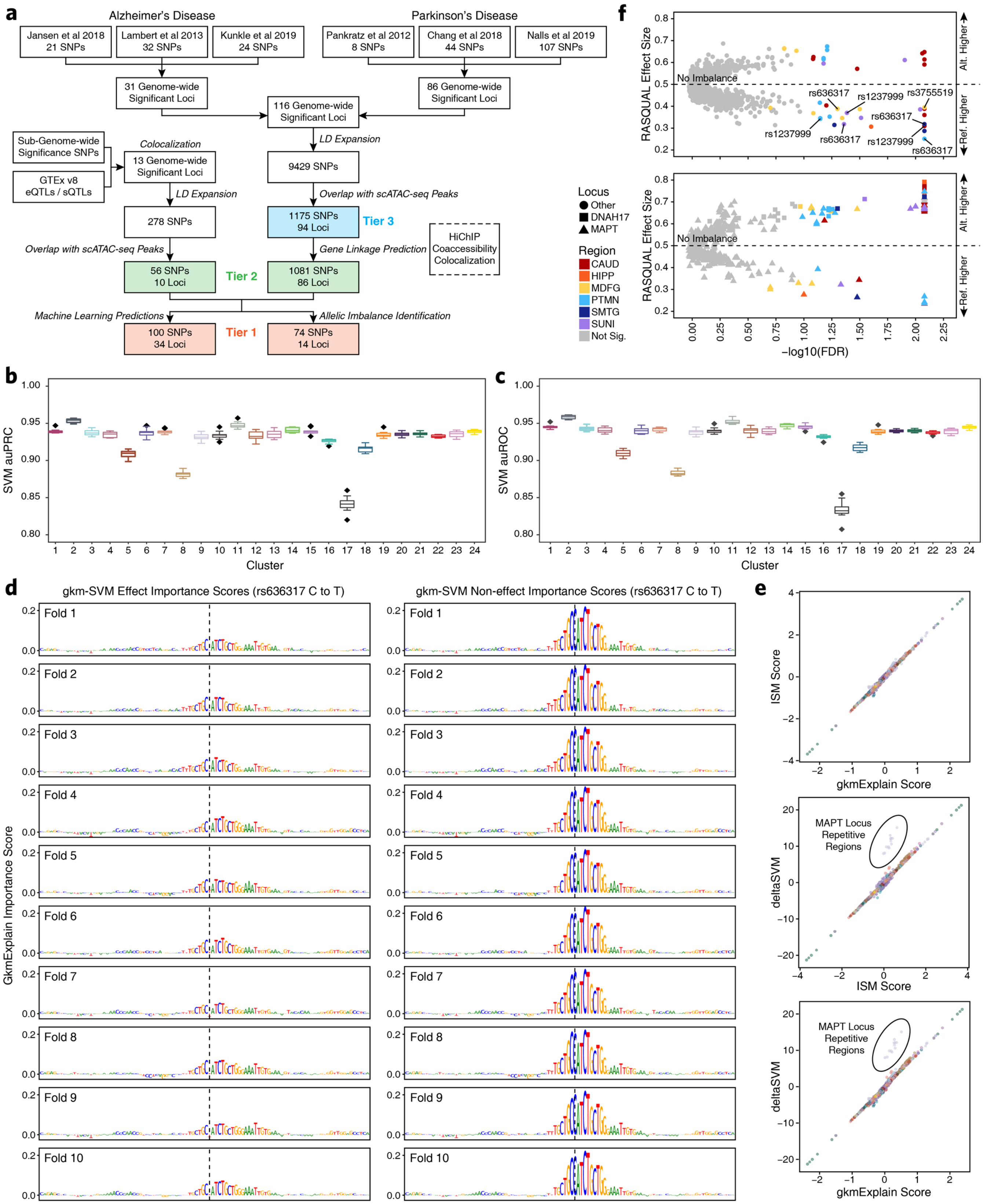
Extended Data Fig. 3 | Neuronal sub-clustering identifies diverse biologically relevant populations of neurons. a–**d**, UMAP dimensionality reduction of neuronal cells (identified as Clusters 1-7, 11, and 12 from Fig. 1e) (**a**) prior to or (**b**–**d**) after batch correction with Harmony of scATAC-seq data from 10 different samples. Each dot represents a single cell (N = 21,116). Dots are colored by (**a**, **b**) the sample of origin, (**c**) the neuronal sub-cluster (repeated from Fig. 2a), or (**d**) its gene activity score for the annotated lineage-defining gene. In (**d**), gene activity scores were imputed using MAGIC. Grey represents the minimum gene activity score while purple represents the maximum gene activity score for the given gene. The minimum and maximum scores are shown in the bottom left of each panel. The gene of interest is shown in the upper right of each panel. **e**, Heatmap of gene activity scores for all neuronal markers used in identifying relevant cell types for neuronal sub-clusters. Color represents the column-wise z-scores for each gene across all neuronal sub-clusters with values thresholded at -2 and +2. Neuronal cluster "major annotation" is shown by color along with a cluster description to the right of the plot. **f**–**h**, The same UMAP dimensionality reduction shown in Extended Data Fig. 3c but cells are colored by (**f**) the major cell class annotation, (**g**) a more granular neuronal sub-annotation, or (**h**) the neuronal cell class annotation. Assignment was made based on gene activity scores of lineage-defining genes. The cell class annotation shown in (**h**) was used to perform LD score regression analysis.



Extended Data Fig. 4 | Sub-clustering of cells from the substantia nigra identifies TH-positive dopaminergic neurons. a–d, UMAP dimensionality reduction after iterative LSI of scATAC-seq data from substantia nigra cells from 2 different samples. Each dot represents a single cell (N = 11,199). Dots are colored by (**a**) their corresponding substantia nigra sub-cluster, (**b**) the sample of origin, or (**c**, **d**) its gene activity score for (**c**) the tyrosine hydoxylase (*TH*) gene, a specific marker of dopaminergic neurons or (**d**) other lineage-defining genes. In (**c**, **d**), gene activity scores were imputed using MAGIC. Grey represents the minimum gene activity score while purple represents the maximum gene activity score. The minimum and maximum scores are shown in the bottom left of each panel. In (**a**-**c**), the predicted cluster cell type identities are overlaid on the UMAPs.

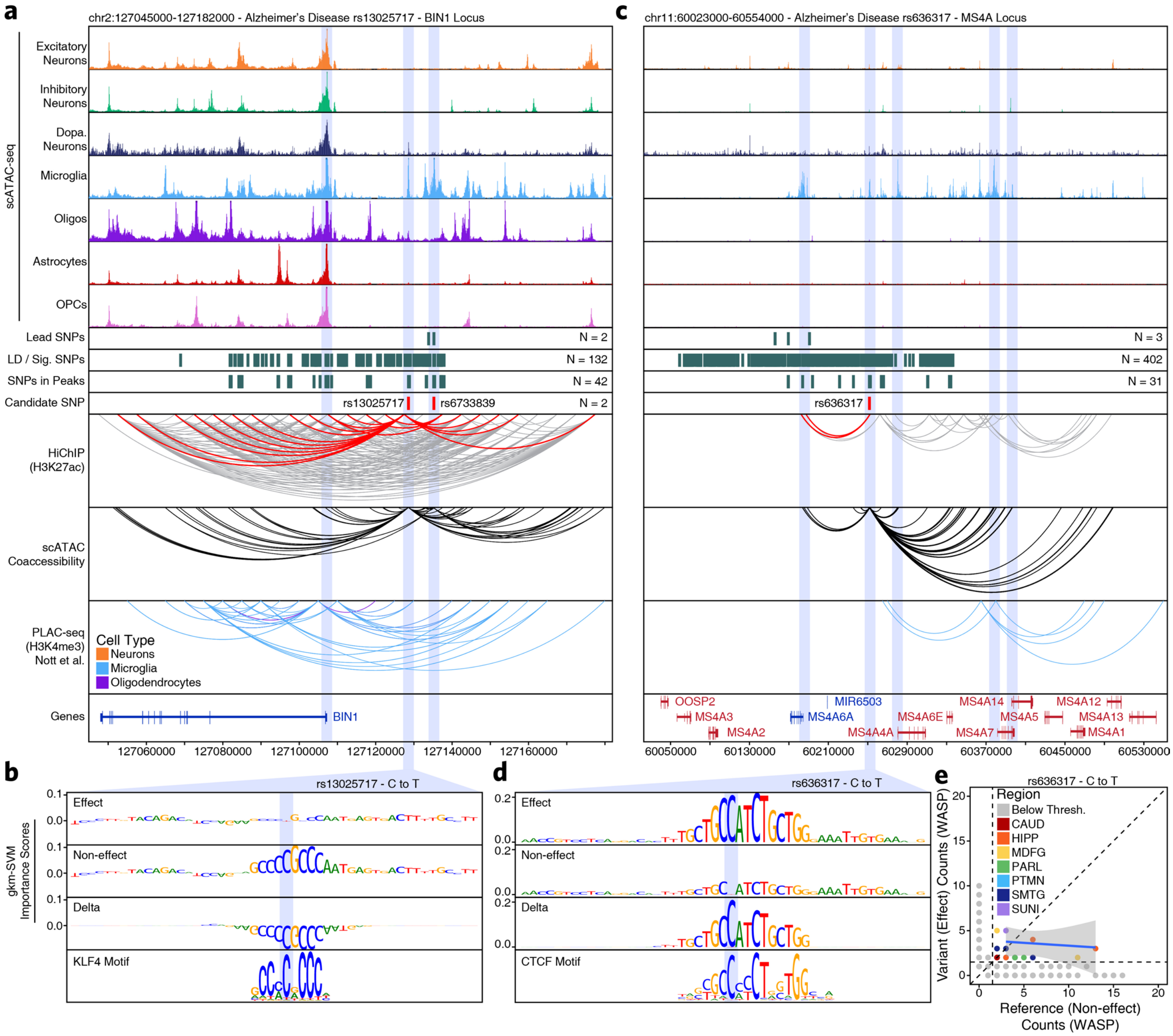


Extended Data Fig. 5 | HiChIP and co-accessibilty predict enhancer-promoter interactions in primary adult human brain. a, Heatmap representation of HiChIP interaction signal at 100-kb, 25-kb, and 5-kb resolution at the *OLIG2* locus. Sample shown represents the substantia nigra from donor 03-41. Signal is normalized to the square root of the coverage. The maximum value of the color range and the coordinates along chromosome 21 are shown below each panel. **b**, Bar plots showing the total number of paired-end reads sequenced for each HiChIP library generated in this study. Color represents the brain region from which the data was generated. **c**, Bar plots showing the number of valid interaction pairs identified in HiChIP data from all samples profiled in this study. Color represents the type of interaction identified. **d**, Bar plot showing the overlap of FitHiChIP loop calls from the 4 gross brain regions profiled. Color indicates whether the loop was identified in a single region (unique) or more than one region (shared). **e**, Bar plot showing the classification of FitHiChIP loop calls based on whether the loop call contained an ATAC-seq peak (from either the bulk ATAC-seq peak set or the scATAC-seq peak set) or TSS in one, both, or no anchor. **f**, Bar plots showing the number of Cicero-predicted co-accessibility-based peak links that are observed in HiChIP (left) or the number of HiChIP-based FitHiChIP loop calls that are predicted as peak links by Cicero. **g**, Bar plot showing the number of cell type-specific peaks (defined as peaks identified through feature binarization; N = 221,062) or non-cell type-specific peaks (defined as scATAC-seq peaks that were not identified through feature binarization; N = 137,960) that overlap or do not overlap a Cicero-predicated co-accessibility linkage. Significance determined by Kolmogorov-Smirnov test.

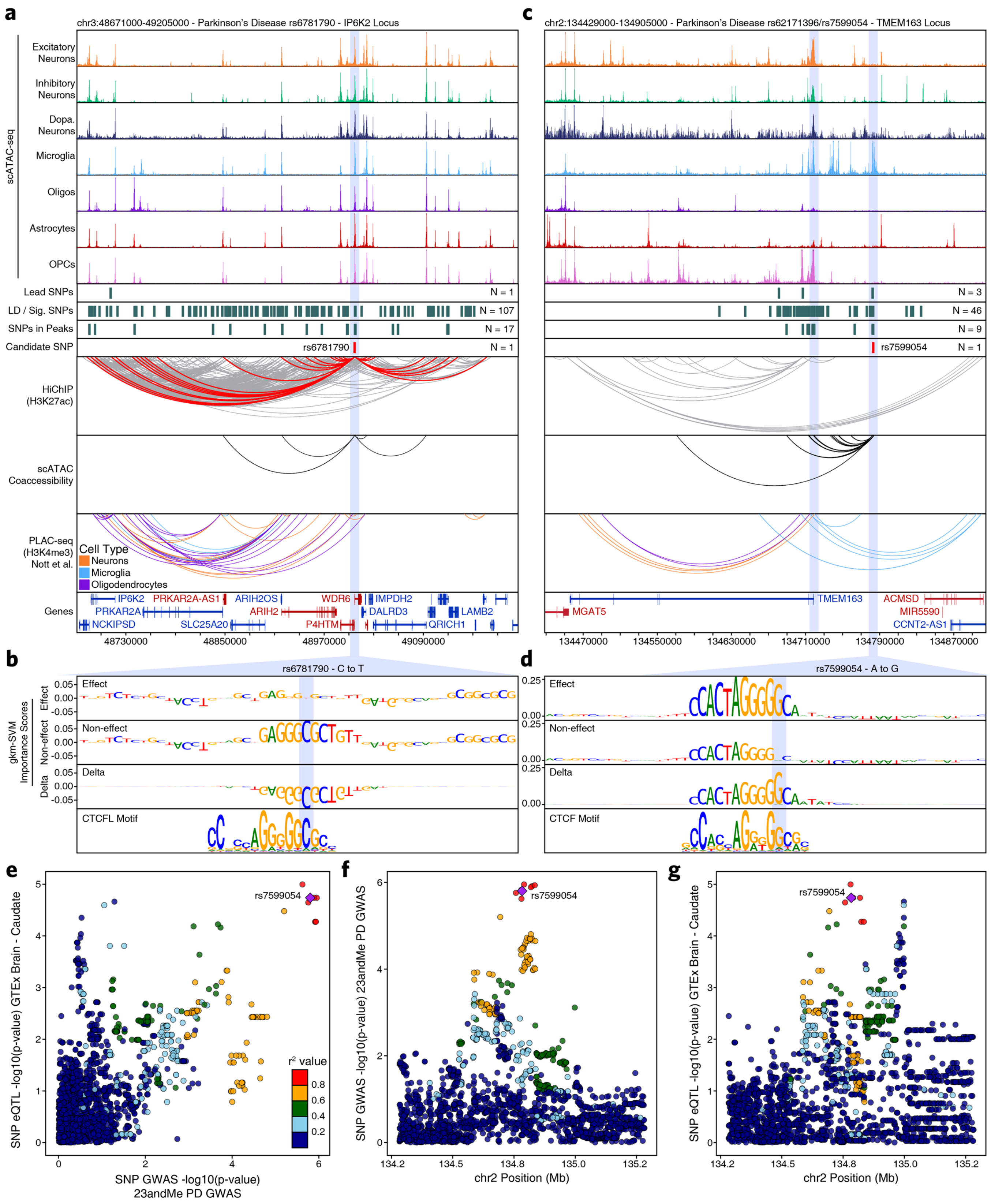


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | A multi-omic tiered approach leverages machine learning to predict functional noncoding SNPs in AD and PD. a, Flow chart of the analytical framework used to prioritize noncoding SNPs and predict functionality. The highest confidence SNPs (Tier 1) are supported by either machine learning predictions, allelic imbalance, or both. Moderate confidence SNPs (Tier 2) are supported by the presence of the SNP within a peak and a HiChIP loop or co-accessibility peak link that connects the SNP to a gene. Lower confidence SNPs (Tier 3) are only supported by the presence of the SNP in a peak. **b**, **c**, Box plot showing the area under (**b**) the precision-recall curve or (**c**) the receiver-operating characteristics curve for the gkm-SVM machine learning classifier. Performance for each of the 24 broad clusters is shown with dots representing outliers. The lower and upper ends of the box represent the 25th and 75th percentiles. The whiskers represent 1.5 multiplied by the inter-quartile range. The center line represents the median. **d**, GkmExplain importance scores shown across all 10 folds for each base across a 100-bp window surrounding rs636317 for the effect (left) and non-effect (right) bases. **e**, Dot plots showing comparison of the GkmExplain score, ISM score, and deltaSVM score. Each dot represents an individual SNP test in a given fold. Dot color represents the GWAS locus number. The only off-diagonal dots (circled) correspond to repetitive regions within the *MAPT* locus where the deltaSVM score appears to be particularly sensitive. **f**, Dot plot showing allelic imbalance assessed by RASQUAL across all bulk ATAC-seq data used in this study from a region-specific analysis. Significance is assessed by RASQUAL (see Methods). Dot color indicates the brain region found to have significant allelic imbalance. Grey dots do not pass significance testing based on an empircal distribution of permuted null q-values and a 10% false discovery rate. A RASQUAL effect size greater than 0.5 indicates that the alternate allele is enriched while less than 0.5 indicates that the reference allele is enriched. The plot is divided to show SNPs within the *MAPT* and *DNAH17* loci (bottom) and SNPs in all other loci (top). SNPs mentioned in downstream analyses are highlighted by red text.

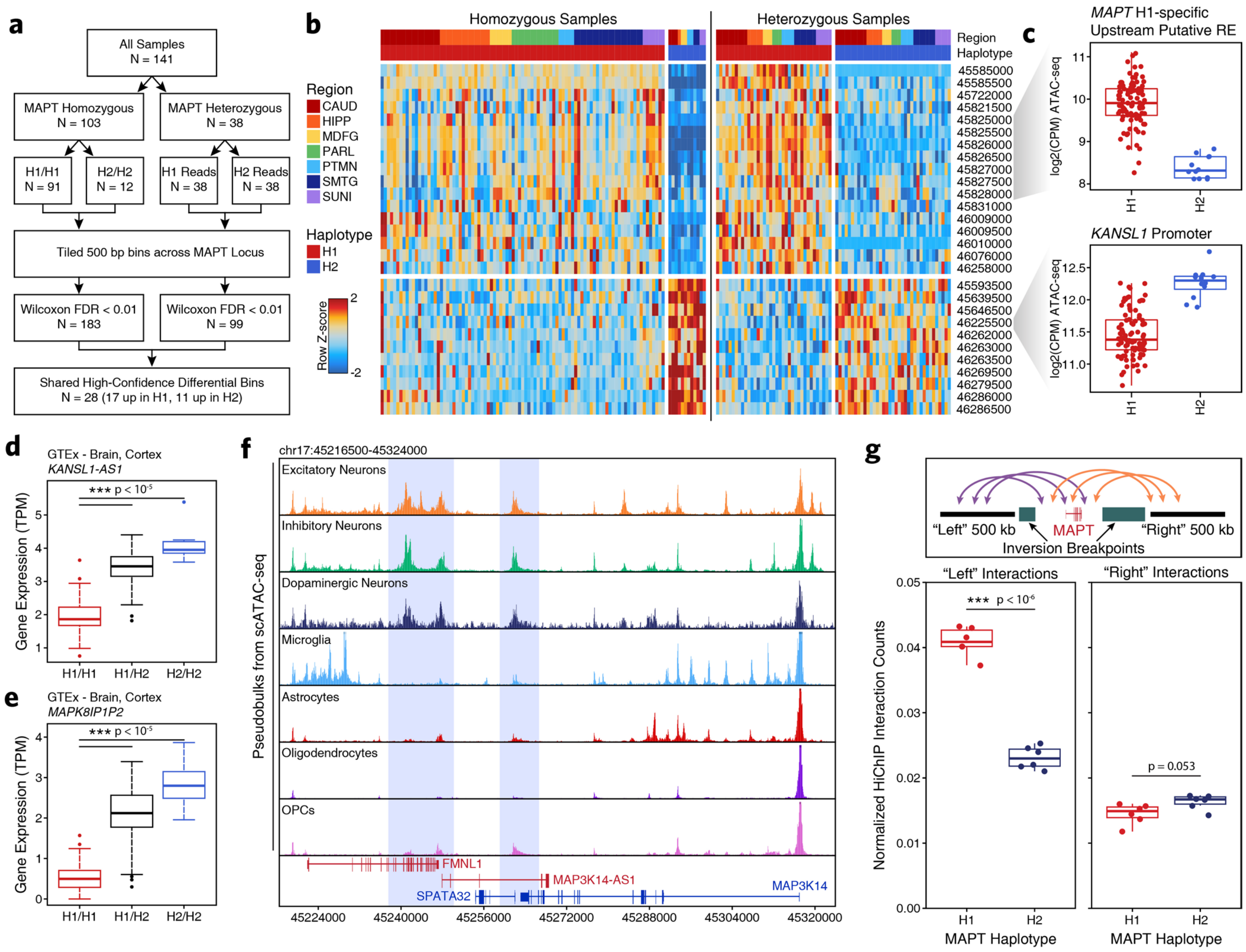


Extended Data Fig. 7 | Multi-omic characterization of well-studied AD-related GWAS loci pinpoints putative functional noncoding SNPs. **a**, **c**, Normalized scATAC-seq-derived pseudo-bulk tracks, H3K27ac HiChIP loop calls, co-accessibility correlations, and publically available H3K4me3 PLAC-seq loop calls (Nott. et al. 2019) in (**a**) the *BIN1* gene locus (chr2:127045000-127182000) and (**c**) the *MS4A* gene locus (chr11:60023000-60554000). scATAC-seq tracks represent the aggregate signal of all cells from the given cell type and have been normalized to the total number of reads in TSS regions, enabling direct comparison of tracks across cell types. For HiChIP, each line represents a FitHiChIP loop call connecting the points on each end. Red lines contain one anchor overlapping the SNP of interest while grey lines do not. For co-accessibility, only interactions involving the accessible chromatin region of interest are shown. For PLAC-seq, MAPS loop calls from microglia (blue), neurons (orange), and oligodendrocytes (purple) are shown. **b**, **d**, GkmExplain importance scores for each base in the 50-bp region surrounding (**b**) rs13025717 or (**d**) rs636317 for the effect and non-effect alleles from the gkm-SVM model for microglia (Cluster 24). The predicted motif affected by the SNP is shown at the bottom and the SNP of interest is highlighted in blue. **e**, Dot plot showing allelic imbalance at rs636317. Significance of allelic imbalance was determined by RASQUAL. The bulk ATAC-seq counts determined by WASP and ASEReadCounter for the reference/non-effect (A) allele and variant/effect (T) allele are plotted. Each dot represents an individual bulk ATAC-seq sample (N = 140) colored by the brain region from which the sample was collected. Samples where fewer than 3 reads were present to support both the reference and variant allele (that is presumed homozygotes or samples with insufficient sequencing depth) are shown in grey. The blue line represents a linear regression of the non-grey points and the grey box represents the 95% confidence interval of that regression.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Multi-omic characterization of noncoding SNPs identifies novel genes implicated in PD. a, **c**, Normalized scATAC-seq-derived pseudo-bulk tracks, H3K27ac HiChIP loop calls, co-accessibility correlations, and publically available H3K4me3 PLAC-seq loop calls (Nott. et al. 2019) in (**a**) the *IP6K2* gene locus (chr3:48671000-49205000) or (**c**) the *TMEM163* gene locus (chr2:134429000-134905000). scATAC-seq tracks represent the aggregate signal of all cells from the given cell type and have been normalized to the total number of reads in TSS regions, enabling direct comparison of tracks across cell types. For HiChIP, each line represents a FitHiChIP loop call connecting the points on each end. Red lines contain one anchor overlapping the SNP of interest while grey lines do not. For co-accessibility, only interactions involving the accessible chromatin region of interest are shown. For PLAC-seq, MAPS loop calls from microglia (blue), neurons (orange), and oligodendrocytes (purple) are shown. **b**, **d**, GkmExplain importance scores for each base in the 50-bp region surrounding (**b**) rs6781790 or (**d**) rs7599054 for the effect and non-effect alleles from the gkm-SVM model for (**b**) astrocytes (Cluster 15) or (**d**) microglia (Cluster 24). The predicted motif affected by the SNP is shown at the bottom and the SNP of interest is highlighted in blue. **e**, Dot plot comparing the –log10(p-value) from 23andMe PD GWAS data with the –log10(p-value) from GTEx Caudate eQTL data of SNPs in the *TMEM163* locus. Each dot represents an individual SNP. Dot color represents the r^2 value of LD with the lead SNP (rs7599054 – purple diamond) within a European reference population. **f**, **g**, Dot plots showing the genomic coordinates of each SNP and the –log10(p-value) from (**f**) 23andMe PD GWAS data or (**g**) GTEx Caudate eQTL data. Dots are colored as in Extended Data Fig. 8e. In (**e**–**g**), p-values are based on genome-wide chi-squared statistics from the relevant GWAS and eQTL studies.



Extended Data Fig. 9 | Epigenomic dissection of the *MAPT* locus. **a**, Flowchart illustrating the analytical scheme used to identify bins with significant allelic imbalance across the H1 and H2 *MAPT* haplotypes. **b**, Heatmaps showing chromatin accessibility in 500-bp bins identified as having significantly different accessibility across *MAPT* haplotypes. Regions are shown for homozygous samples without allelic read splitting (left) and for heterozygous samples after allelic read splitting (right). Bin start coordinates are shown to the right. **c**, Box and whiskers plots for multiple regions which show differential chromatin accessibility across the H1 and H2 *MAPT* haplotypes. Each dot represents a single homozygous H1 (N = 91) or homozygous H2 (N = 12) sample. Heterozygotes are not shown. The lower and upper ends of the box represent the 25th and 75th percentiles. The whiskers represent 1.5 multiplied by the inter-quartile range. The center line represents the median. **d**, **e**, Gene expression of (**d**) the *KANSL1-AS1* gene or (**e**) the *MAPK8IP1P2* gene shown as a box plot from GTEx cortex brain samples subdivided based on *MAPT* haplotype. The lower and upper ends of the box represent the 25th and 75th percentiles. The whiskers represent 1.5 multiplied by the inter-quartile range. The center line represents the median. ***$p < 10^{-5}$ based on Wilcoxon rank sum test. N = 117 H1/H1, 78 H1/H2, and 10 H2/H2. **f**, Sequencing tracks from pseudo-bulk data derived from predicted cell types in scATAC-seq data. This region represents a zoomed in view of the predicted distal regulatory region (chr17:45216500-45324000) that interacts with the *MAPT* promoter in the H1 haplotype. Putative neuron-specific regulatory elements are highlighted in blue. **g**, Box plots showing differential HiChIP interaction signal occurring between regions within the *MAPT* inversion and regions outside the inversion ("left" or "right"). The schematic at the top explains the analysis performed. The box plots show normalized HiChIP interaction counts for the H1 (N = 6) and H2 (N = 6) haplotypes for upstream/"left" interactions and downstream/"right" interactions. P-value determined by paired two-sided t-test.

nature research

Corresponding author(s): Howard Y. Chang and Thomas J. Montine

Last updated by author(s): Aug 11, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

□ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection
Data analysis	All custom code used in this work is available in the following GitHub repository: https://github.com/kundajelab/alzheimers_parkinsons. For publicly available software, the following versions apply: MACS2 v2.2.7.1; ArchR 0.9.4; bcftools (v1.7-1.9), HiC-Pro v2.11.0; Seurat v2.3; FitHiChIP v7; bedtools 2.26.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated in this work is available through GEO accession GSE147672. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147672

To facilitate broad access to our data, we have created WashU Epigenome browser session (Session ID: drS3o1n4kJ) for our scATAC-seq data in the following track formats: (i) broad cell types (“Corces_scATAC_BroadCellTypes”), (ii) broad clusters (“Corces_scATAC_BroadClusters”), (iii) neuron subclusters (“Corces_scATAC_NeuronSubClusters”), and (iv) neuron subclustered cell types / LDSC groups (“Corces_scATAC_NeuronSubCellTypes”). These tracks are accessible via the following link - http://epigenomegateway.wustl.edu/legacy/?genome=hg38&session=drS3o1n4kJ.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample size was based on the available biological material. However, sufficient samples were used to allow statistical measures of reproducibility across multiple biological donors in all cases.

Data exclusions

Sequencing data that did not pass pre-established quality control filters was excluded from analysis. For bulk ATAC-seq, samples with low signal-to-noise were uniformly excluded to prevent misinterpretation. For scATAC-seq and HiChIP, all generated data passed quality control filters. All conclusions were validated through replication.

Replication

Replication across biological samples was the primary metric for reproducibility of sequencing data. For our machine learning work, 10-fold cross validation was used.

Randomization

During nuclei isolation, tissue samples were randomized into batches to avoid batch effects from nuclei isolation. During bulk ATAC-seq, scATAC-seq, and HiChIP library construction, randomized batches were also used.

Blinding

All brain tissue nuclei isolation and bulk ATAC-seq data generation was carried out in a blinded manner. HiChIP and scATAC-seq were performed after nuclei were isolated, bulk ATAC-seq was performed, and the data was un-blinded. Thus, blinding was not possible for HiChIP, or scATAC-seq but these techniques were performed on nuclei isolated in a blinded fashion as mentioned above.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Post-mortem human brain samples were used in this work. These samples were taken from individuals ranging from 38 to 93 years old (mean of 80.8) from primarily Caucasian ancestry. Some of these individuals were cognitively assessed. No individuals had mutations for known drivers of Alzheimer's or Parkinson's diseases. Donors were 41% female, 59% male. Additional donor characteristics are available in Supplementary Table 1.

Recruitment

Participants were research volunteers in the Stanford, Arizona, or University of Washington Alzheimer's Disease Research Center, or the Stanford Morris K. Udall Center of Excellence for Parkinson's Disease Research, who consented to donate their brains for research following each institutions IRB-approved protocol. Bias may exist due to regional or socioeconomic factors that constrain the patient populations at these facilities which could bias towards an over-representation of caucasian individuals. No other self-selection biases are expected.

Ethics oversight

Post-mortem brain samples were collected with approved consent and overseen by the relevant institutional review boards of Stanford University, the University of Washington, and Banner Health.

Note that full information on the approval of the study protocol must also be provided in the manuscript.