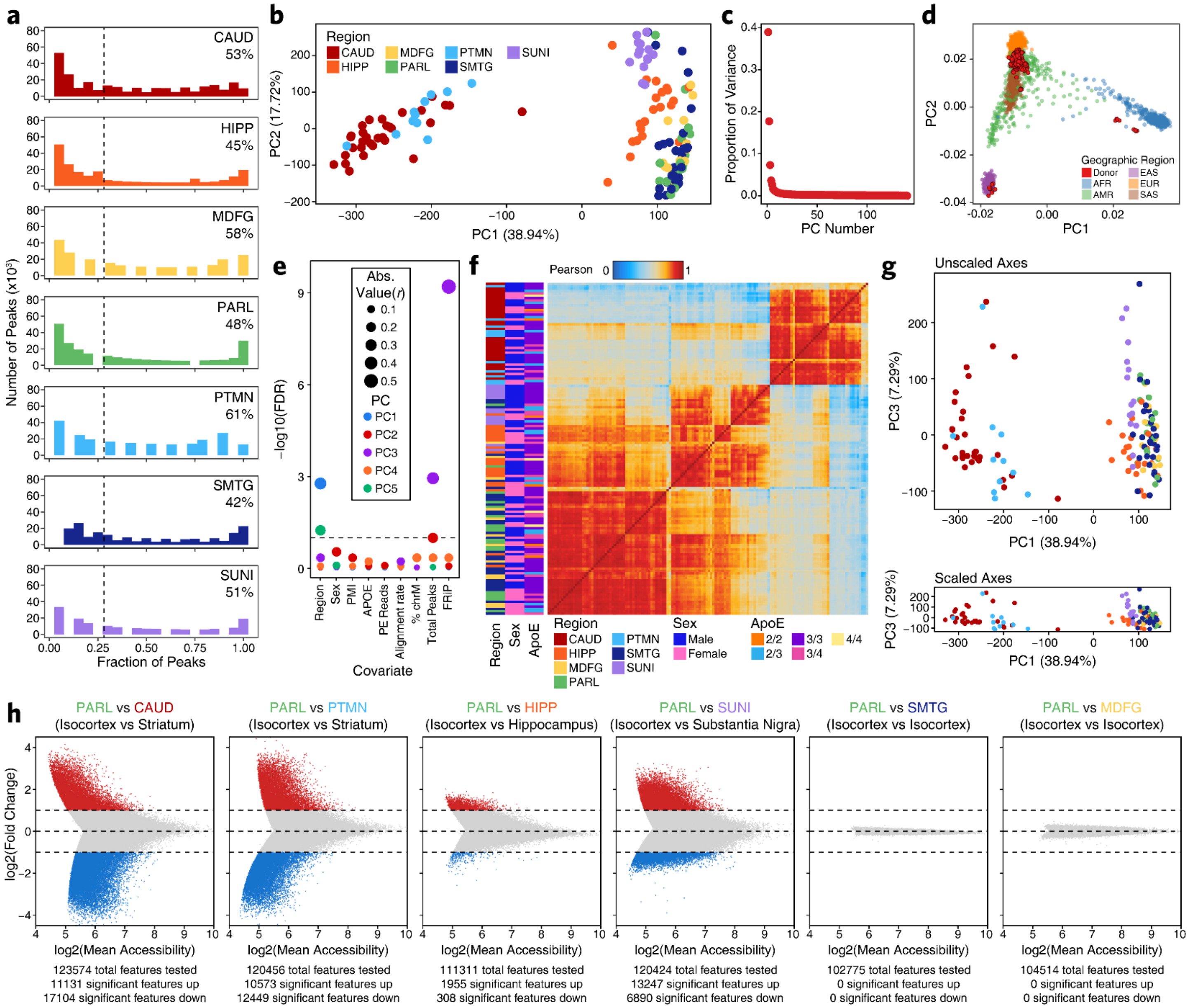
SUPPLEMENTARY INFORMATION FOR:

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

M. Ryan Corces, Anna Shcherbina, Soumya Kundu, Michael J. Gloudemans, Laure Frésard, Jeffrey M. Granja, Bryan H. Louie, Tiffany Eulalio, Shadi Shams, S. Tansu Bagdatli, Maxwell R. Mumbach, Boxiang Liu, Kathleen S. Montine, William J. Greenleaf, Anshul Kundaje, Stephen B. Montgomery, Howard Y. Chang*, Thomas J. Montine*

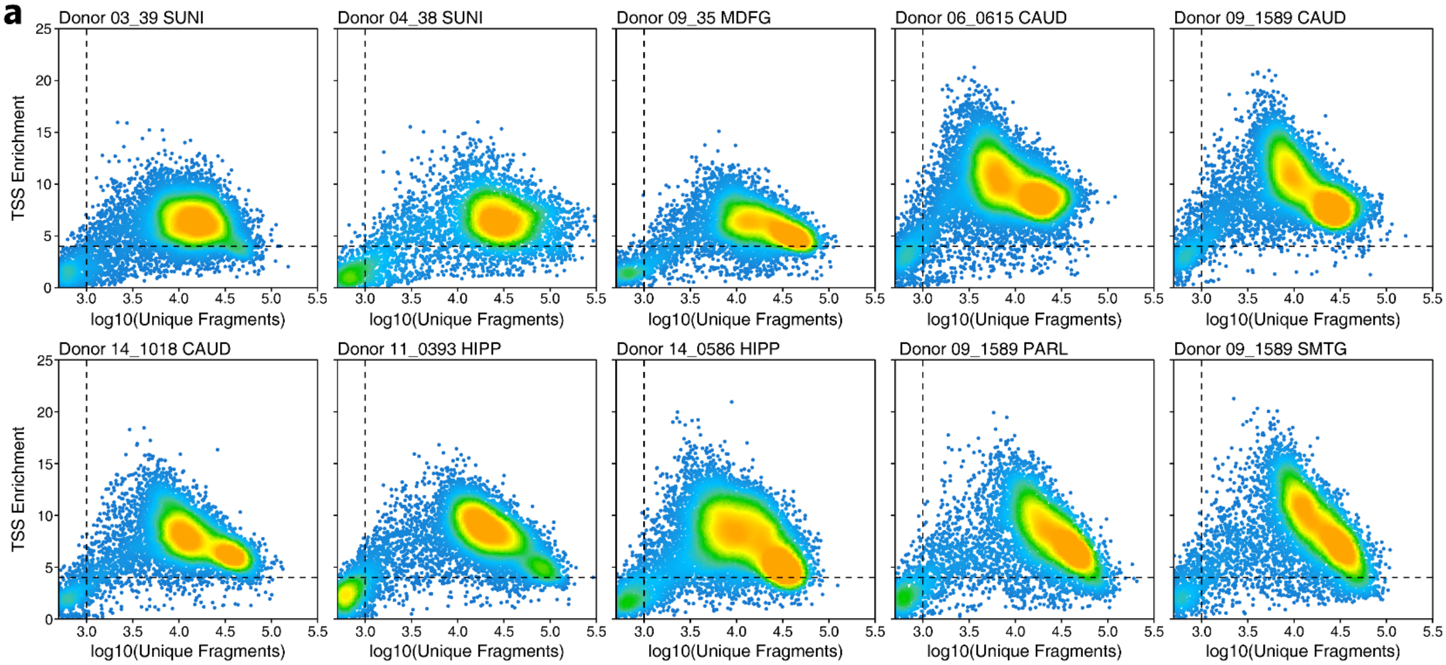
*Correspondence should be addressed to T.J.M. (tmontine@stanford.edu) or H.Y.C. (howchang@stanford.edu)

This document contains the following:



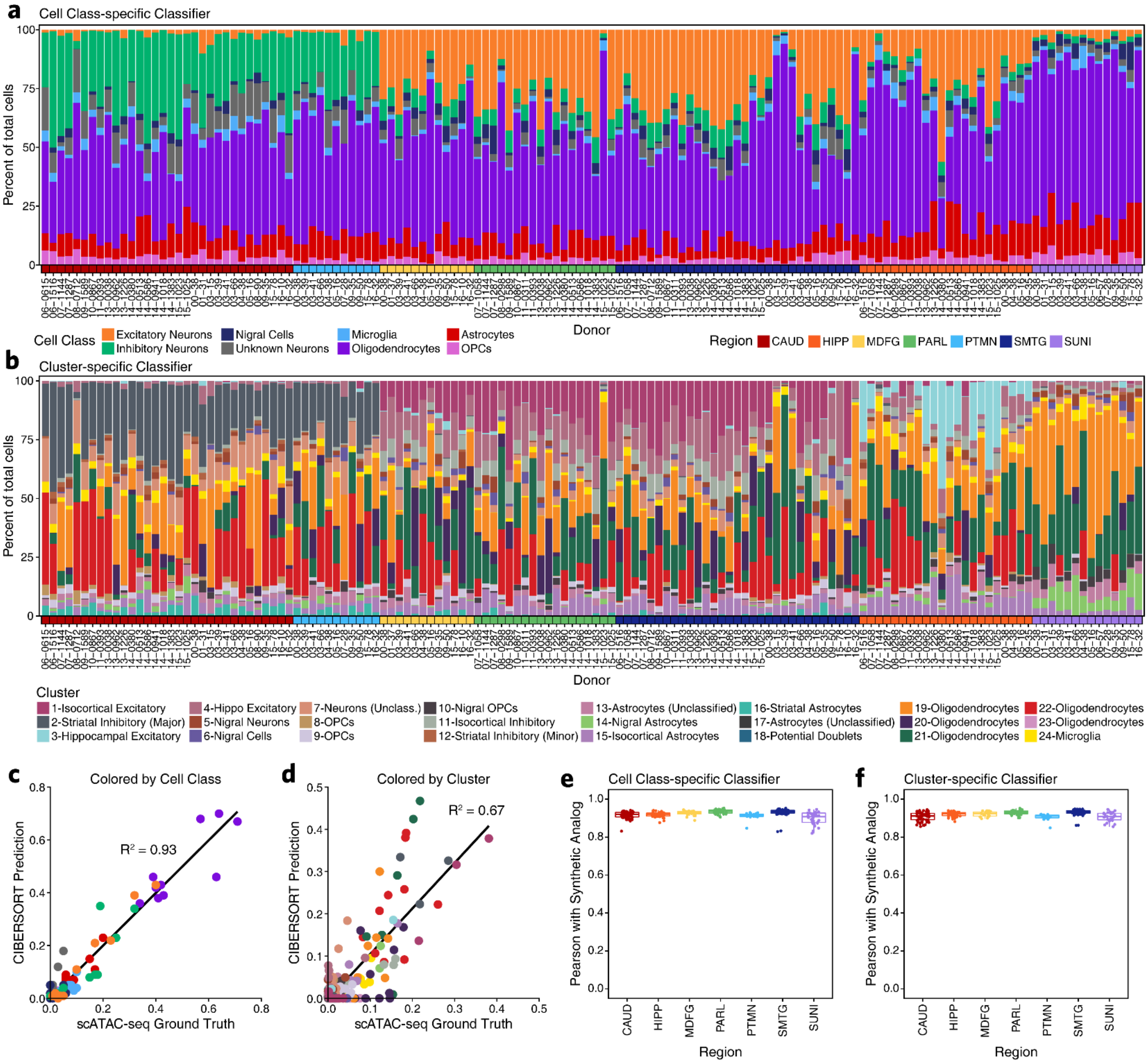
Supplementary Figure 1 - Analysis of bulk ATAC-seq data from adult brain identifies brain-regional heterogeneity.

a, Bar plots of peak reproducibility across all bulk ATAC-seq biological replicates from the designated brain region. In each panel, the dotted line represents the cutoff for the fraction of samples that must have called a peak for the peak to be included in our merged reproducible bulk ATAC-seq peak set (cutoff = 0.3 or 30%). For each region, the percent of the total peaks passing this cutoff is indicated in the upper right. **b**, Principal component analysis of all samples showing components 1 and 2. Each dot represents a single piece of tissue with technical replicates merged where applicable. Color represents the brain region from which the sample was isolated. The proportion of variance explained is included along each axis and the axes have been scaled to these values. **c**, Dot plot showing the proportion of variance explained by each principal component in the analysis presented in Supplementary Figure 1b. **d**, Principal component analysis of genotypes from the 1000 genomes project showing different races depicted by color. Biological samples from the current study (in red) were genotyped and projected into the principal component space defined by the 1000 genomes data to confirm the self-reported race of all individuals. **e**, Dot plot showing the significance of correlation between covariates and each of the top 5 principal components. Significance of each covariate is determined by the p-value of its contribution to a linear model. Dot size represents the absolute value of the correlation while color represents the principal component number. Covariates included were: region, biological sex, post-mortem interval (PMI), *APOE* genotype, paired-end (PE) reads, cumulative alignment rate, percent of reads mapping to chrM (%chrM), the total peaks called by MACS2 per sample, and the fraction of reads in peaks (FRiP) within those peaks. **f**, Sample by sample Pearson correlation heatmap of all 140 samples profiled in this study. Brain region, donor biological sex, and *APOE* genotype are indicated by color to the left. **g**, Principal component analysis of all samples showing components 1 and 3. Each dot represents a single piece of tissue with technical replicates merged where applicable. Color represents the brain region from which the sample was isolated. The proportion of variance explained is included along each axis. In the top panel, the axes have not been scaled to the proportion of variance explained to enable visualization of the distribution along PC3, which is correlated with data quality). In the bottom panel, the axes have been scaled to the proportion of variance explained to contextualize the relevance of PC3 to the overall data. **h**, MA plots showing the change in normalized bulk ATAC-seq accessibility comparing the parietal lobe (PARL) to all other brain regions. Each dot represents an individual peak from the merged bulk ATAC-seq peak set. Only peaks that showed non-zero accessibility in at least one sample were tested for significance.



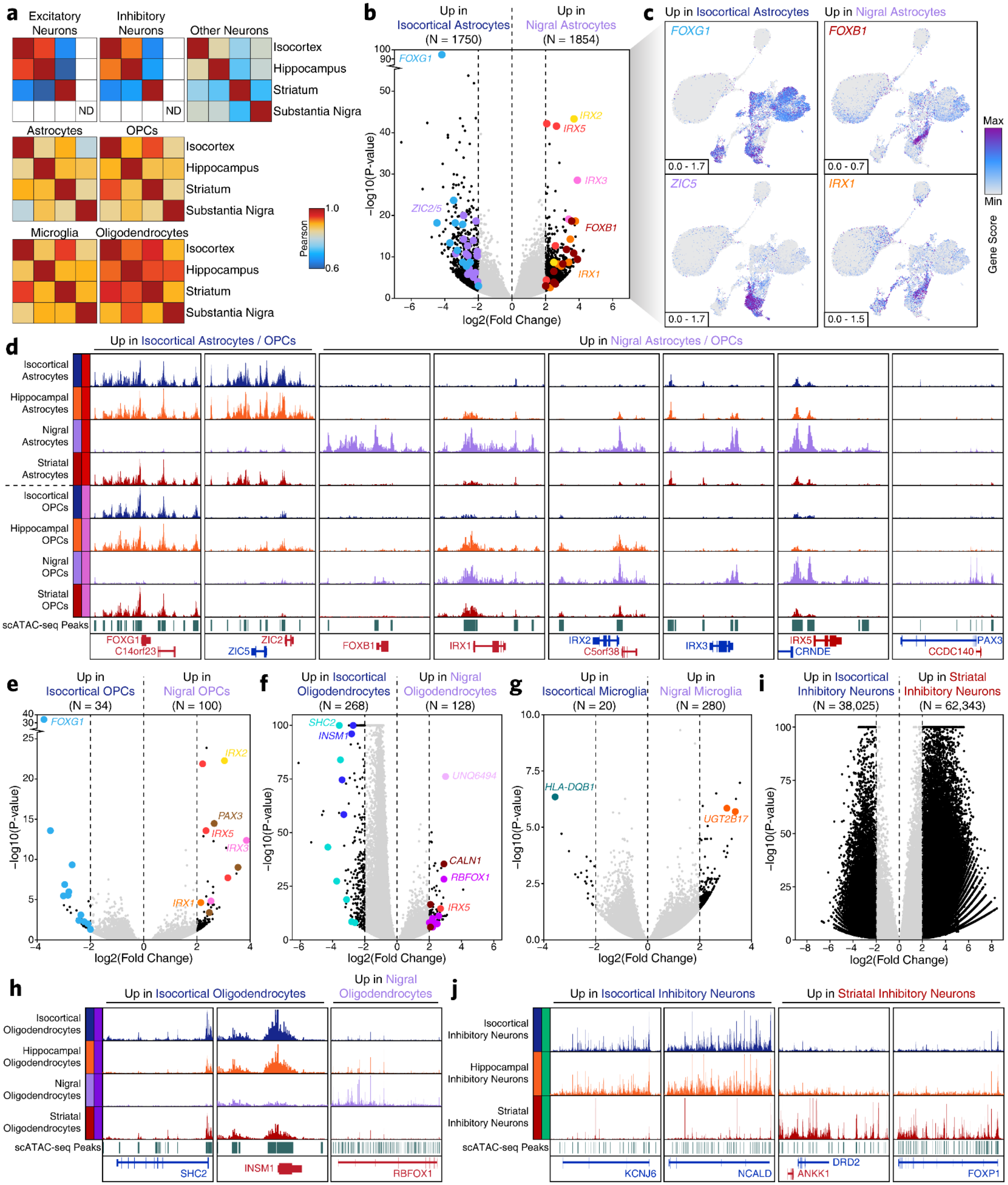
Supplementary Figure 2 – Quality control of scATAC-seq libraries.

a, Dot plots showing the TSS enrichment score and the total number of fragments per cell for each of the 10 samples profiled by scATAC-seq. Each dot represents an individual cell. Dot color represents density on the plot. Dotted lines represent the quality control cutoffs implemented.



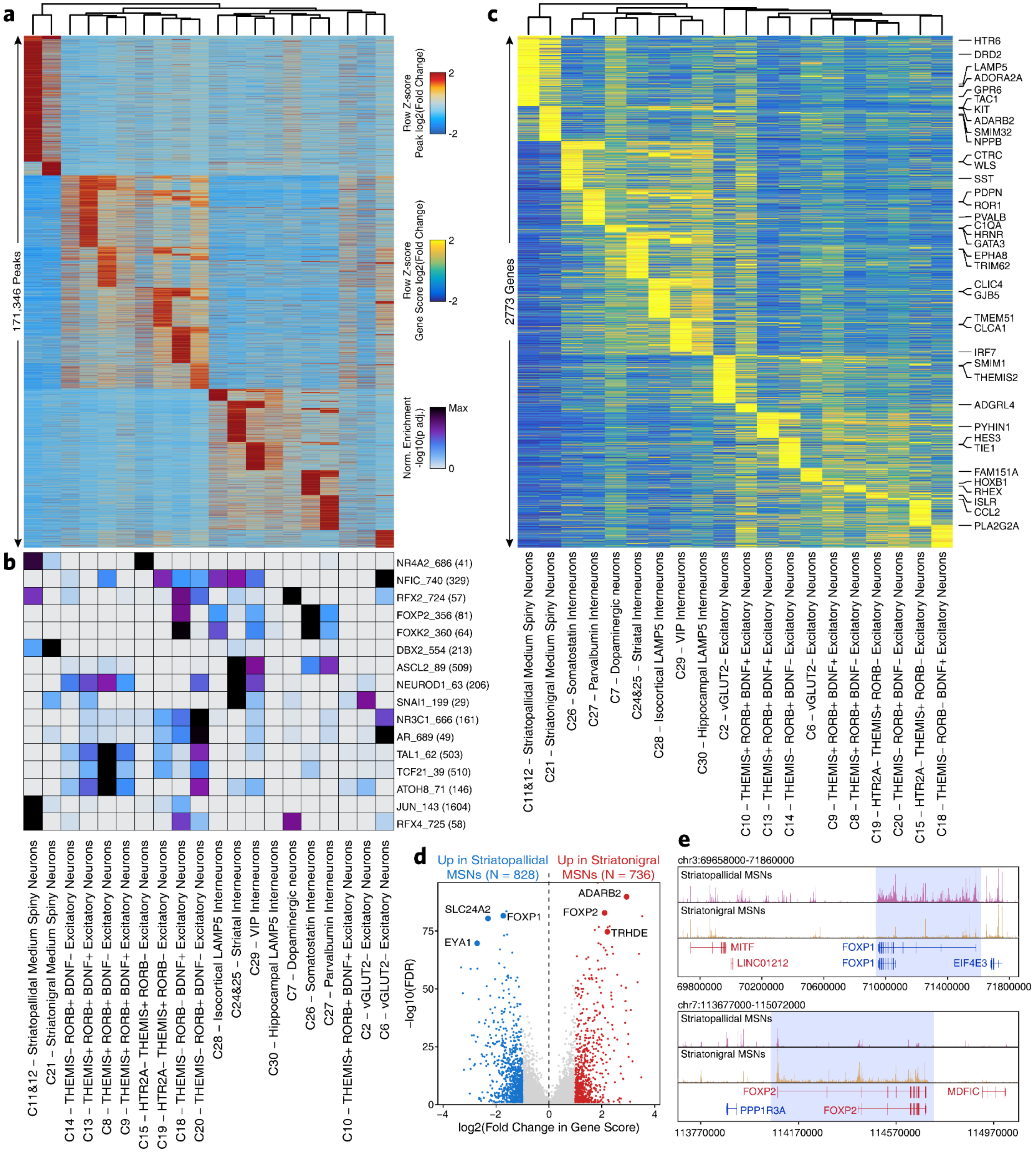
Supplementary Figure 3 - Cell type-specific scATAC-seq data enables deconvolution of chromatin accessibility data from bulk regions in the adult brain.

a-b, Bar plot showing CIBERSORT predictions across all bulk ATAC-seq data generated in this study. Samples are sorted and colored (bottom of plot) by the region from which they were profiled. Bars are colored by (**a**) the predicted cell class or (**b**) the predicted cluster. Donor IDs are annotated below the plot. **c-d**, Dot plot showing the performance of the CIBERSORT classifier by comparing the “ground truth” from scATAC-seq data and the CIBERSORT prediction on the bulk ATAC-seq data from the same tissue sample. Each dot represents (**c**) a cell class (i.e. the merge of multiple clusters) or (**d**) a cluster from one of the 10 scATAC-seq samples profiled. Dots are colored by (**c**) cell class according to the legend in Supplementary Figure 3a or (**d**) cluster according to the legend in Supplementary Figure 3b. **e-f**, Box and whiskers plot showing the Pearson correlation of each bulk ATAC-seq sample with a synthetic analog derived from admixing (**e**) the proportional cell class-specific signal predicted by the cell class-specific CIBERSORT classifier or (**f**) the proportional cluster-specific signal predicted by the cluster-specific CIBERSORT classifier (see Methods). 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. All samples are shown as individual dots overlaid on the box and whiskers (N = 59 CAUD, 42 HIPP, 23 MDFG, 35 PARL, 20 PTMN, 61 SMTG, 28 SUNI).


a
Excitatory Neurons
Inhibitory Neurons
Other Neurons
ND
ND
Isocortex
Hippocampus
Striatum
Substantia Nigra
Astrocytes
OPCs
Isocortex
Hippocampus
Striatum
Substantia Nigra
Microglia
Oligodendrocytes
Isocortex
Hippocampus
Striatum
Substantia Nigra
Pearson
1.0
0.6
b
Up in Isocortical Astrocytes (N = 1750)
Up in Nigral Astrocytes (N = 1854)
FOXG1
IRX2
IRX5
IRX3
ZIC2/5
FOXB1
IRX1
−log10(P-value)
log2(Fold Change)
c
Up in Isocortical Astrocytes
Up in Nigral Astrocytes
FOXG1
FOXB1
ZIC5
IRX1
0.0 - 1.7
0.0 - 0.7
0.0 - 1.7
0.0 - 1.5
Max
Min
Gene Score
d
Up in Isocortical Astrocytes / OPCs
Up in Nigral Astrocytes / OPCs
Isocortical Astrocytes
Hippocampal Astrocytes
Nigral Astrocytes
Striatal Astrocytes
Isocortical OPCs
Hippocampal OPCs
Nigral OPCs
Striatal OPCs
scATAC-seq Peaks
FOXG1
C14orf23
ZIC2
ZIC5
FOXB1
IRX1
IRX2
C5orf38
IRX3
IRX5
CRNDE
PAX3
CCDC140
e
Up in Isocortical OPCs (N = 34)
Up in Nigral OPCs (N = 100)
FOXG1
IRX2
PAX3
IRX5
IRX3
IRX1
−log10(P-value)
log2(Fold Change)
f
Up in Isocortical Oligodendrocytes (N = 268)
Up in Nigral Oligodendrocytes (N = 128)
SHC2
INSM1
UNQ6494
CALN1
RBFOX1
IRX5
−log10(P-value)
log2(Fold Change)
g
Up in Isocortical Microglia (N = 20)
Up in Nigral Microglia (N = 280)
HLA-DQB1
UGT2B17
−log10(P-value)
log2(Fold Change)
i
Up in Isocortical Inhibitory Neurons (N = 38,025)
Up in Striatal Inhibitory Neurons (N = 62,343)
−log10(P-value)
log2(Fold Change)
h
Up in Isocortical Oligodendrocytes
Up in Nigral Oligodendrocytes
Isocortical Oligodendrocytes
Hippocampal Oligodendrocytes
Nigral Oligodendrocytes
Striatal Oligodendrocytes
scATAC-seq Peaks
SHC2
INSM1
RBFOX1
j
Up in Isocortical Inhibitory Neurons
Up in Striatal Inhibitory Neurons
Isocortical Inhibitory Neurons
Hippocampal Inhibitory Neurons
Striatal Inhibitory Neurons
scATAC-seq Peaks
KCNJ6
NCALD
DRD2
ANKK1
FOXP1

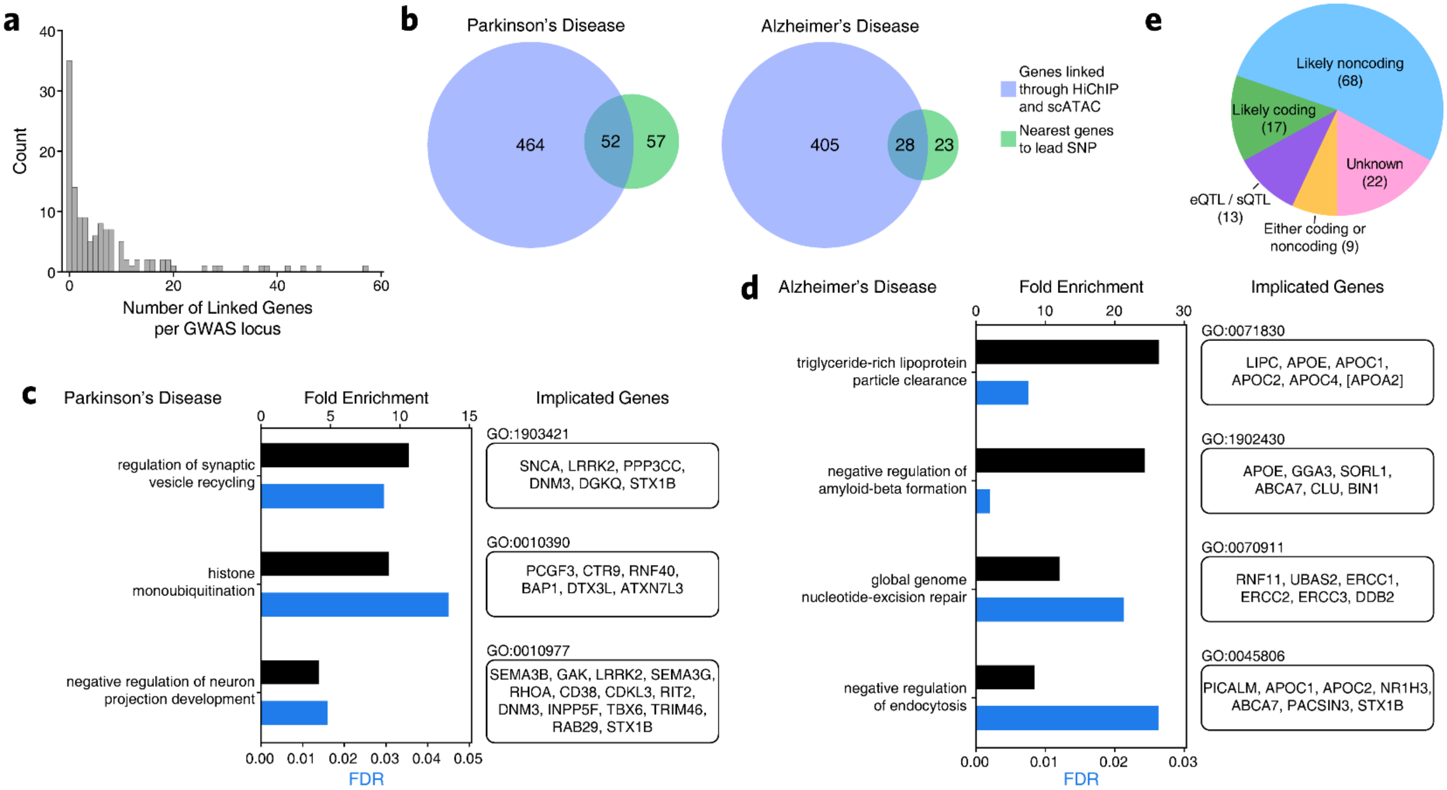
Supplementary Figure 4 - scATAC-seq reveals epigenetic encoding of region-specific cellular gene regulatory programs

a, Pearson correlation heatmaps showing the correlation of pseudo-bulk replicates from cell types across different brain regions. Cell type signals were generated by making at least 2 non-overlapping pseudo-bulk replicates of at least 150 cells. Cases where insufficient cells were present to make these pseudo-bulk replicates were excluded from analysis (ND) to avoid overinterpretation. All heatmaps use the same color scale ranging from R values of 0.6 to 1.0. **b**, Volcano plot of peaks that show differential signal between astrocytes from the substantia nigra and astrocytes from the isocortex. Peaks below a log2(fold change) threshold of 2 were not considered. Peaks near genes that are predicted to be key lineage-defining genes are accented with larger colored dots. Significance determined by the Wald test (DESeq2). **c**, The same UMAP dimensionality reduction shown in Figure 1e but each cell is colored by its gene activity score for the annotated lineage-defining gene identified in Supplementary Fig. 4b (*FOXG1*, *ZIC5*, *FOXB1*, *IRX1*). 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 left of each panel. **d**, Sequencing tracks of pseudo-bulk scATAC-seq data from multiple genomic regions showing differential chromatin accessibility between astrocytes or OPCs in the isocortex and substantia nigra. From left to right: Isocortex-specific - *FOXG1* (chr14:28750000-28787000), and *ZIC2*/*ZIC5* (chr13:99937000-99999000); Substantia Nigra-specific - *FOXB1* (chr15:59996000-60012000), *IRX1* (chr5:3589600-3607800), *IRX2* (chr5:2737000-2760000), *IRX3* (chr16:54277000-54292000), *IRX5* (chr16:54927000-54940000), and *PAX3* (chr2:222189500-222333500). Peaks called in scATAC-seq data are shown below each plot. Each pseudo-bulk track was derived from merging all single cells corresponding to the annotated cell types in the specified regions. Tracks have been normalized to the total number of reads in TSS regions, enabling direct comparison of tracks within each vertical panel. **e-g**, Volcano plot of peaks that show differential signal between (**e**) OPCs, (**f**) oligodendrocytes, or (**g**) microglia from the substantia nigra and isocortex. Peaks below a log2(fold change) threshold of 2 were not considered. Peaks near genes that are predicted to be key lineage-defining genes are accented with larger colored dots. Significance determined by the Wald test (DESeq2). **h,j**, Sequencing tracks of pseudo-bulk scATAC-seq data from multiple genomic regions showing differential chromatin accessibility between (**h**) oligodendrocytes from the substantia nigra and isocortex or (**j**) inhibitory neurons from the striatum and isocortex. From left to right: (**h**) Isocortex-specific - *SHC2* (chr19:409800-463200), and *INSM1* (chr20:20361000-20374000); Substantia nigra-specific - *RBFOX1* (chr16:5899200-7791000); (**j**) Isocortex-specific - *KCNJ6* (chr21:37583000-37955000), and *NCALD* (chr8:101673000-102141000); Striatum-specific - *DRD2* (chr11:113369000-113602000), and *FOXP1* (chr3:70922000-71622000). Peaks called in scATAC-seq data are shown below each plot. Each pseudo-bulk track was derived from merging all single cells corresponding to the annotated cell types in the specified regions. Tracks have been normalized to the total number of reads in TSS regions, enabling direct comparison of tracks within each vertical panel. **i**, Same as Supplementary Fig. 4e but for inhibitory neurons in the isocortex and striatum. Significance determined by the Wald test (DESeq2).



Supplementary Figure 5 – Neuronal cell class-specific peaks and genes delineate differences between biologically relevant neuronal cell types

a-c, Heatmap of neuronal cell class-specific (**a**) peaks, (**b**) corresponding TF motifs, or (**c**) gene activity scores identified by ArchR. Each row represents a different (**a**) peak region (N = 171,346), (**b**) TF motif or (**c**) gene. For **a** and **c**, color represents the row-wise Z-score of the log2(Fold Change) between the given neuronal cell class and a background set of features with color values thresholded at 2 and -2. For **b**, color represents the normalized enrichment (-log10(p-value) of the hypergeometric test) of the TF motif in the relevant peaks with color values thresholded at the maximum enrichment of that motif as indicated in parentheses next to the TF name. **d**, Volcano plot of differential gene activity scores between striatopallidal and striatonigral medium spiny neurons (MSNs). Select genes of interest are labeled and highlighted by larger dots. Significance determined by the Wald test (DESeq2). **e**, Sequencing tracks of pseudo-bulk scATAC-seq data from striatopallidal MSNs and striatonigral MSNs at the *FOXP1* locus (chr3:69658000-71860000) and the *FOXP2* locus (chr7:113677000-115072000). Each pseudo-bulk track was derived from merging all single cells corresponding to the annotated cell types in the specified regions. Tracks have been normalized to the total number of reads in TSS regions, enabling direct comparison of tracks within each vertical panel.



Supplementary Figure 6 - HiChIP and co-accessibility implicates disease-relevant genes in AD and PD through linkage of noncoding GWAS SNPs to target genes

a, Histogram of the number of genes linked per GWAS locus. Each bar represents a bin of size 1. A link represents a putative association of a SNP within an ATAC-seq peak to a gene based on HiChIP or co-accessibility data. **b**, Venn diagram of the number of genes linked through assessment of the nearest gene to the lead SNP (green) or the number of genes linked though HiChIP and scATAC-seq analyses of LD-expanded polymorphisms (blue) for all GWAS loci from AD (left) and PD (right). **c-d**, GO-term enrichments of genes linked to (**c**) AD and (**d**) PD GWAS SNPs through FitHiChIP loop calls or scATAC-seq based co-accessibility correlations. Significance determined by the hypergeometric test. **e**, Characterization of GWAS loci in AD and PD according to the predicted effects of the polymorphisms. For example, loci whose phenotypic association is likely mediated by changes in coding regions are marked as "Likely coding". Loci where no known SNPs overlap an exonic region are annotated as "Likely noncoding". Loci whose effect could be mediated by either coding or noncoding mechanisms are marked as "Either coding or noncoding" whereas loci with no polymorphisms overlapping a peak region from our bulk or scATAC-seq data or an exonic region are marked as "Unknown".

SUPPLEMENTARY NOTES

Supplementary Note 1 - Quality control analysis of bulk ATAC-seq data

In total, we generated 268 bulk ATAC-seq libraries from 140 macrodissected brain samples, with technical replicates for 128 of the 140 samples. As with t-SNE (Figure 1c), principal component analysis (PCA) shows clear brain region-specific clustering with 39% of the variance explaining the difference between striatal and non-striatal brain regions (Supplementary Fig. 1b-c). For example, region-specific chromatin accessibility was observed at the dopamine receptor D2 (*DRD2*) gene in the striatum, corresponding to medium spiny neurons[1], the Iroquois homeobox 3 (*IRX3*) gene in the substantia nigra, corresponding to diencephalic-origin astrocytes[2], or the potassium voltage-gated channel modifier subfamily S member 1 (*KCNS1*) gene in the isocortex, corresponding to various neuronal populations[3]. To validate the reported race of each donor, we performed genotype PCA using the 1000 genomes data as a reference (Supplementary Fig. 1d and Supplementary Table 1). These samples showed no clustering based on covariates such as post-mortem interval, *APOE* genotype, or biological sex (Supplementary Fig. 1e-f), and we identified very few peaks with significant differential accessibility across sexes (Supplementary Data Set 1). We note that PC3, representing 7.29% of variance, shows a significant correlation with metrics of data quality including the fraction of reads in peaks (Supplementary Fig. 1e,g and Supplementary Table 1). Lastly, we note that comparison of bulk ATAC-seq data across regions from the same anatomical tissue type (i.e. isocortex) showed no significant differences (Supplementary Fig. 1h).

Supplementary Note 2 - Single-cell ATAC-seq provides reference cell populations for deconvolution of cell type-specific signals in bulk data

Using the cell type-specific signals present in our scATAC-seq data, we performed cell type deconvolution of our bulk ATAC-seq data using CIBERSORT[4]. From our 8 cell classes and our 24 clusters, we created classifiers to deconvolve the ATAC-seq signal from all 140 samples profiled by bulk ATAC-seq in this study (Supplementary Fig. 3a-b and Supplementary Data Set 5). These classifiers recapitulate expected patterns of cell type abundance such as a relative absence of excitatory neurons in the striatum (Supplementary Fig. 3a) and mapping of signal from Cluster 14 (nigral astrocytes) specifically to samples from the substantia nigra (Supplementary Fig. 3b). Moreover, these classifiers predict a wide range of cellular composition across the macrodissected human brain samples used here, even within a single region. Such large differences in cell type composition can hamper efforts to find differential features, further supporting the use of single-cell approaches to understand complex tissues and disease states where small disease-specific variation may be overshadowed by larger differences in cell type composition across samples.

By comparing the CIBERSORT prediction to the observed "ground truth" in the scATAC-seq data for the 10 samples profiled here, we assessed the performance of the cell class-specific and cluster-specific classifiers (Supplementary Fig. 3c-d). As would be expected, the cell type-specific classifier showed better performance than the cluster-specific classifier, largely due to over- or under-prediction of closely related clusters, such as the oligodendrocytic Clusters 19-23, by the cluster-specific classifier (Supplementary Fig. 3d). To benchmark the ability of these classifiers to explain the majority of signal observed in bulk ATAC-seq, we created synthetic analogs[5] by proportionally mixing signal from each cell group (class or cluster) at each peak. For each bulk ATAC-seq sample, we asked how similar this sample is to the synthetic analog. In almost all cases, the Pearson correlation value between each sample and its synthetic analog surpassed 0.9, indicating that the vast majority of bulk ATAC-seq signal can be explained by either the class-specific or cluster-specific classifiers (Supplementary Fig. 3e-f).

Supplementary Note 3 - Single-cell ATAC-seq identifies brain region-specific differences in glial cells

Our dissection of the cell type-specific chromatin landscapes in adult brain identified clusters that are both region- and cell type-specific, such as Cluster 14 which is comprised almost exclusively of astrocytes from the substantia nigra (Extended Data Fig. 1e and Supplementary Data Set 2). This observation indicates that certain brain cell types may show region-specific variation. This phenomenon has been very well described in neurons, with, for example, inhibitory neurons from the striatum (largely medium spiny neurons) differing substantially from inhibitory neurons outside of the striatum[6]. Murine oligodendrocytes[7] and astrocytes[8] also show regional differences in morphology, function, and gene expression. However, the brain-regional variation of glial cells in humans remains less well understood. To address this, we grouped cells into one of the 8 broad cell classes defined above and created region-specific pseudo-bulk profiles from the cumulative data (see Methods). Using these region-cell type combinations, we calculated Pearson correlations for all regions across a single cell type (Supplementary Fig. 4a). As expected, neuronal cell types showed the most regional variation, with a minimum Pearson correlation R value of 0.6.

Glial cells, however, also showed appreciable regional variation, with astrocytes showing the most variation followed by OPCs (Supplementary Fig. 4a). Within astrocytes, the greatest difference was found between the substantia nigra and the isocortex, indicating that the function or composition of astrocytes may differ across these brain regions. Differential peak analysis identified significant differences in chromatin accessibility near transcriptional regulators that may help explain the observed regional astrocytic differences (Supplementary Fig. 4b and Supplementary Data Set 6). In particular, nigral astrocytes showed significantly increased accessibility at the forkhead box B1 (*FOXB1*), *IRX1*, *IRX2*, *IRX3*, and *IRX5* genes. Conversely, isocortical astrocytes showed significantly increased accessibility at the *FOXG1*, zic family member 2 (*ZIC2*), and *ZIC5* genes. These changes in chromatin accessibility were associated with differential motif accessibility (Supplementary Data Set 6) and would be expected to correlate with similar changes in gene expression for the annotated genes. Moreover, the gene activity scores of these genes are definitional for the region-cell subtypes with, for example, *FOXB1* being active only in nigral astrocytes and *ZIC2* and *ZIC5* being active in all other astrocytes (Supplementary Fig. 4c-d). Of particular interest, the observed FOX switch from *FOXG1* in isocortical (and hippocampal/striatal) astrocytes to *FOXB1* in nigral astrocytes and the significant changes in chromatin accessibility at the IRX genes represent a potential transcriptional lineage control mechanism that could help to better understand region-specific functional differences in these astrocytes. Notably, diencephalic brain regions such as the substantia nigra have previously been shown to express *FOXB1*[9], *IRX1*[10], and *IRX3*[2] during early brain development, thus explaining part of this broad TF-based lineage control. These transcriptional regulators could be exploited to drive differentiation programs to, for example, create regionally biased glial cells in vitro.

In addition to controlling regional astrocytic identity, chromatin accessibility at *IRX* genes was also found to differentiate nigral OPCs from isocortical OPCs (Supplementary Fig. 4d-e). Similarly, *FOXG1* also showed significantly more accessibility in isocortical OPCs, echoing the observations from astrocytes. Lastly, chromatin accessibility at the *PAX3* gene locus was significantly higher in nigral OPCs compared to isocortical OPCs (Supplementary Fig. 4d-e). Taken together, these results identify shared and disparate transcriptional regulatory programs that likely control regional differences amongst astrocytes and OPCs in the substantia nigra and isocortex.

Compared to astrocytes, oligodendrocytes and microglia showed less regional variation in chromatin accessibility (Supplementary Fig. 4f-g). While a small number of genes showed highly significant regional differences in oligodendrocytes (Supplementary Fig. 4h), very few genes showed appreciable regional differences among microglia. As noted previously, the regional differences observed in glial cells are a small fraction of the size and magnitude of regional differences observed in neurons (Supplementary Fig. 4i-j), further emphasizing the importance of single-cell approaches to study complex tissues.

Supplementary Note 4 - Single-cell ATAC-seq identifies neuronal cell class-specific biology

We identify peak regions and corresponding transcription factor motifs that are unique to each neuronal cell class, highlighting potential gene regulatory mechanisms underlying the class-specific differences (Supplementary Fig. 5a-b and Supplementary Data Set 7). Additionally, using gene activity scores, we identify genes that show neuronal class-specific activity (Supplementary Fig. 5c), including genes that differentiate striatopallidal and striatonigral medium spiny neurons such as the transcription factors *FOXP1* and *FOXP2* (Supplementary Fig. 5d-e and Supplementary Data Set 7), which have previously been shown to exhibit variable expression in the striatum[11].

Supplementary Note 5 - Tiered approach to identification of functional GWAS polymorphisms

Of the 9,707 putative disease relevant SNPs, 9,429 were included in downstream analysis based on genome-wide significance or presence in high LD with a genome-wide significant SNP. Of these, 1175 SNPs overlapped peak regions identified in the cluster-specific peaks of our single-cell ATAC-seq data (Tier 3). Intersecting these SNPs with gene linkage predictions based on HiChIP, co-accessibility, or colocalization, we identified 1081 SNPs that met the requirements of Tier 2. Additionally, 278 SNPs (of the original 9,707) were included based on colocalization or presence in high LD with a colocalized SNP, despite the SNP of interest not meeting genome-wide GWAS *p*-value thresholds. Of these colocalization-based SNPs, 56 overlapped peak regions from our cluster-specific scATAC-seq data and were therefore also included in Tier 2. Collectively, these merged Tier 2 SNPs implicate 516 and 433 genes putatively affected by the activity of GWAS polymorphisms in PD and AD, respectively (Supplementary Fig. 6a-b). These gene sets are enriched for biological processes known to be implicated in AD and PD including lipoprotein particle clearance[12] (AD) and synaptic vesicle recycling[13] (PD) (Supplementary Fig. 6c-d). Lastly, we identified high-confidence Tier 1 SNPs as the subset of Tier 2 SNPs that were predicted to affect transcription factor binding based on our machine learning framework (100 SNPs) or our allelic imbalance analysis using RASQUAL (74 SNPs) (Extended Data Fig. 6a and Supplementary Table 2).

Supplementary Note 6 - Additional examples of putative functional SNPs identified through multi-omic integrative analysis

In addition to the examples highlighted in the main figures, we identify multiple Tier 1 examples of putative functional noncoding SNPs. In the case of the *BIN1* locus, our work and previous work[14] predict SNP rs6733839 to disrupt a putative MEF2 binding site in a microglia-specific regulatory element located 28-kb upstream of the *BIN1* promoter (Extended Data Fig. 7a). Our machine learning framework additionally implicates SNP rs13025717 which we predict to disrupt a KLF4 binding motif in a microglia-specific putative regulatory element 21-kb upstream of *BIN1* (Extended Data Fig. 7b). Both of these SNPs have previously been shown to have sequence-specific correlations with *BIN1* gene expression[15]. Similarly, we identified rs636317 in the *MS4A6A* locus which shows significant allelic imbalance and disrupts a microglia-specific CTCF binding motif (Extended

Data Fig. 7c-e and Supplementary Data Set 10). Cumulatively, these results annotate the most likely functional SNPs mediating known disease associations in AD and PD (Supplementary Table 2). Importantly, these predicted functional SNPs do not always affect the expected cell type nor target the closest gene, further emphasizing the utility of our integrative multi-omic approach.

Though many such anecdotes exist (Supplementary Table 2), we also noted a pattern whereby many SNPs appear to disrupt binding sites related to the CCCTC-Binding Factor (*CTCF*) protein. For example, SNP rs6781790 disrupts a predicted CTCFL binding site within the promoter of the WD Repeat Domain 6 (*WDR6*) gene (Extended Data Fig. 8a-b). Similarly, SNP rs7599054 disrupts a putative microglia-specific CTCF binding site near the Transmembrane Protein 163 (*TMEM163*) gene (Extended Data Fig. 8c-d). Colocalization analysis of rs7599054 predicts a significant effect of this SNP on *TMEM163* gene expression (Extended Data Fig. 8e-g), providing orthogonal validation of the predicted SNP-to-gene linkage.

Supplementary Note 7 - A multi-omic epigenetic dissection of the MAPT gene locus

The *MAPT* gene encodes tau isoforms, primarily neuronal microtubule binding proteins that, under pathologic conditions, can adopt an abnormal structure and extensive post translational modifications, a process called neurofibrillary degeneration, which is a hallmark of AD and other neurodegenerative diseases, but not PD[16]. Enigmatically, *MAPT* is a replicated risk locus for PD despite the absence of neurofibrillary degeneration[17,18]. The *MAPT* locus, found on chromosome 17, represents one of the largest LD blocks in the human genome (1.8 Mb) and is present in two distinct haplotypes, H1 and H2, the latter formed by an approximately 900 kb inversion of H1 that occurred about 3 million years ago and is present mostly in Europeans[19]. Cumulatively, previous work supports *MAPT* haplotype-specific impacts on transcript amount, transcript stability, and alternative splicing in several neurodegenerative disorders[20–22]. We highlight multiple epigenetic avenues through which the *MAPT* gene is differentially regulated in the H1 and H2 haplotypes, thus explaining at least a portion of the molecular underpinnings of the observed *MAPT* GWAS association in PD.

SUPPLEMENTARY METHODS

Ancestry determination via PCA analysis on genomic data

Genotyping was performed on the bulk ATAC-seq datasets with the bcftools (1.7)[23]. The `bcftools mpileup` command was executed on individual bulk ATAC-seq filtered bam files to generate read pileups. The output of this command was fed into `bcftools call` to perform variant calling on the mpileup files. The resulting vcf files were merged with `bcftools merge`, converted to plink 1.9[24] –bfile format, and filtered to include variants with population minor allele frequency (MAF) greater than or equal to 0.05. Chromosome 1 data from phase 3 of the 1000 Genomes Project[25] was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/GRCh38_positions/ALL.chr1_GRCh38.genotypes.20170504.vcf.gz. The variants were filtered to those with MAF > 0.05. Common variants were identified from the 1000 Genomes SNP set and the donor SNP set above, and the datasets were merged into a single PLINK binary format (bed) file. This yielded 447,096 SNPs for 2916 individuals in the combined 1000 Genomes / donor dataset. The PLINK –pca command was executed with the subject id's of the 1000 Genomes Project individuals passed to the `–family –pca-cluster-names` flags to ensure that PCA would be performed on the 1000 Genomes cohort, and the unknown donors from this study would be projected on the resulting PC's. Individuals form 1000 Genomes and the donors from this study were jointly plotted along PC1 and PC2, and the ancestry for all donors was set as the ancestry of the closest 1000 Genomes individual in PC space.

SNP selection for colocalization testing

A single test for colocalization of GWAS and eQTL association signals involves a locus, a GWAS, an eQTL tissue, and a gene expressed in that tissue. For each GWAS, we selected the set of all loci for which the lead GWAS variant had p-value < 1e-5. Using eQTLs from GTEx brain tissues in the GTEx v8 dataset, we then found all tissue-gene combinations for which the lead SNP at one of the GWAS loci had an eQTL SNP (association p-value < 1e-5) for that gene in that GTEx tissue. This resulted in a list of unique combinations of GWAS trait / genomic locus / eQTL tissue / eQTL gene, each to be tested individually for colocalization of GWAS and eQTL signals. The GWAS threshold of 1e-5 is less stringent than the threshold for genome-wide significance, but we favored sensitivity over specificity when selecting which SNPs to test, since colocalization with a strong eQTL signal may still suggest that a sub-threshold GWAS locus has an expression-mediated effect on disease.

Colocalization analysis

For each colocalization test combination as defined above, we selected all 1000 Genomes Phase 3 variants within a window of 500kb around the lead GWAS variant. We narrowed this list down to SNPs measured not only in the 1000 Genomes VCF, but also in the GWAS and eQTL summary statistics for the selected trait, tissue, and gene. We used a streamlined version of the FINEMAP tool[26] to compute posterior causal probabilities for each SNP at the locus in both the GWAS and eQTL studies, and then combined these probabilities as described in eCAVIAR[27] to compute a colocalization posterior probability (CLPP) score for this test locus. We considered a SNP weakly colocalized if its CLPP score exceeded 0.01; although this seems like a low probability, we have observed previously that loci exceeding this cutoff show considerable likelihood of haring causal eQTL and GWAS variants[28], and our goal in this analysis was to be as sensitive as possible in selecting putatively functional loci for subsequent orthogonal analysis steps.

CIBERSORT deconvolution

CIBERSORT[4] was used to deconvolve bulk ATAC-seq data using signature matrices generated from scATAC-seq data. Default parameters were used. For the cell type-specific classifier, pseudo-bulk replicates were generated for each of the 8 main cell types. For the cluster-specific classifier, pseudo-bulk replicates were generated for each of the 24 clusters.

gkm-SVM machine learning classifier training and testing

For each of the 24 scATAC-seq clusters, we used a 10 fold cross-validation scheme to train weighted gapped k-mer Support Vector Machine (gkm-SVM) models to classify 1000 bp sequences into two classes - accessible (corresponding to sequences underlying peaks) and inaccessible (GC matched inaccessible genomic regions). The test sets for each of the 10 folds are as follows. Fold 0 consisted of chr 1. Fold 1 consisted of chr 2 and chr 19. Fold 2 consisted of chr 3 and chr 20. Fold 3 consisted of chr 6, chr 13, and chr 22. Fold 4 consisted of chr 5, chr 16, and chr Y. Fold 5 consisted of chr 4, chr 15, and chr 21. Fold 6 consisted of chr 7, chr 14, and chr 18. Fold 7 consisted of chr 11, chr 17, and chr X. Fold 8 consisted of chr 9 and chr 12. Fold 9 consisted of chr 8 and chr 10.

For each of the 24 scATAC-seq clusters, we merged the IDR peaks with identical genomic coordinates (peaks with multiple summits) while preserving the summit position and the MACS2 p-value of the peak with the lowest p-value among the ones with the identical coordinates. Next, we ranked the peaks by the MACS2 p-value, expanded each peak by 500 bp on either side of the summit, to a total of 1000 bp, and eliminated those peaks with any ‘N’ bases in the 1000 bp. For each of 10 cross-validation folds, we kept up to 60,000 of the top peaks belonging to the training set and all of the peaks belonging to the much smaller test set, all of which comprised the positively labeled (accessible) examples for training.

In order to generate the negative (inaccessible) examples for each of the cross-validation folds in each single-cell cluster, first, we used seqdataloader (https://github.com/kundajelab/seqdataloader) to generate all 1000 bp sequences obtained by tiling the hg38 genome 200 bp at a time, with a stride of 50 bp, keeping those 200 bp segments that have no IDR peak summits in that cluster, and then expanding those 200 bp segments by 400 bp on each side for a total of 1000 bp. Next, we calculated the GC content of the selected positive examples and all other bins in the genome. We partitioned the positive examples into 20 equally numerous GC bins according to the GC-content percentile of the positive sequence with respect to the positive set. We assigned sequences from all other bins in the genome to GC bins according to their GC-content. Starting with an empty negative set, we then sampled a positive example, sampled a negative sequence from the same GC bin as the sampled positive example, added the negative sequence to the negative set, and repeated this process until the number of negative examples equaled the number of positive examples for both the training set and the test set.

For each of the 10 folds in each of the 24 clusters, we used the 1000-bp DNA sequences corresponding to the positive and GC-matched negative training examples as inputs to the gkmtrain function from the LS-GKM package[29] with the default options, producing a total of 240 models; the default options for LS-GKM included the gapped *k*-mer + center weighted (wgkm) kernel (t = 4), a word length of 11 (l = 11), 7 informative columns (k = 7), 3 maximum mismatches to consider (d = 3), an initial value of the exponential decay function of 50 (M = 50), a half-life parameter of 50 (H = 50), a regularization parameter of 1.0 (c = 1.0), and a precision parameter of 0.001 (e = 0.001). We used the resulting support vectors for each trained model to score the DNA sequences corresponding to the positive and GC-matched negative test set examples for each fold in each cluster by running gkmpredict, and used the scikit-learn python library[30] to calculate both auROC and auPRC accuracy metrics.

gkm-SVM allelic scores of candidate SNPs

We intersected the coordinates of all LD-expanded candidate AD and PD GWAS and colocalization SNPs with those of the peaks for each single-cell ATAC-seq cluster to obtain the SNPs in each cluster that are in peaks. For each SNP in a peak in each of the clusters, we retrieved the 1000 bp DNA sequence around the SNP, with the SNP at its center, and created a sequence corresponding to the effect allele by replacing the 500th position of the sequence with the effect allele. Similarly, we created another sequence corresponding to the non-effect allele by replacing the 500th position of the sequence with the non-effect allele. Furthermore, we repeated the same procedure to also produce 50 bp sequences for each SNP with the effect allele and the non-effect allele by retrieving the 50 bp DNA sequence around each SNP and replacing the 25th position with the effect and the non-effect allele, respectively.

For each SNP in a peak in each of the clusters, we computed **GkmExplain**[31] importance scores for each position in each of the 1000 bp effect and non-effect allele sequences using each of the 10 gkm-SVM[32] models for the respective cluster. GkmExplain is a method to infer the importance or predictive contribution of every base in an input sequence to its corresponding output prediction from a gkm-SVM model. Next, for each SNP in a given cluster, we computed the average score for each position across all 10 models (from the 10 folds) for that cluster for both the effect allele sequence and the non-effect allele sequence, producing a set of consensus importance scores for both the effect allele and the non-effect allele. Then, we subtracted the sum of these consensus importance scores corresponding to the central 50 bp of the non-effect allele sequence from that of the effect allele sequence to compute the GkmExplain score for each SNP in each cluster.

To compute ***in silico*** **mutagenesis (ISM)** scores for each SNP in a peak in each of the clusters, we used each of the 10 fold gkm-SVM models from the respective cluster to compute model output prediction scores for the 50 bp effect and non-effect allele sequences by running gkmpredict. Then, we subtracted the score of the non-effect allele sequence from the score of the effect allele sequence to obtain the ISM score and computed the average ISM score for each SNP across all 10 folds in each cluster.

To compute **deltaSVM** scores, we generated all possible non-redundant k-mers of size 11 and scored each of them using each of the 240 models. Next, for each SNP in a peak in each of the clusters, we used each of the 10 sets of *k*-mer scores from the 10-fold gkm-SVM models from the respective cluster to run deltaSVM[33] on the 50 bp effect and non-effect allele sequences. We computed the average of the resulting deltaSVM scores for each SNP across all 10 folds in each cluster.

Statistical significance and high confidence sets of gkm-SVM based allelic scores for candidate SNPs

In order to obtain a statistical significance for each of the three gkm-SVM model based allelic SNP scores (GkmExplain, ISM and deltaSVM), for each SNP scored in each cluster, we generated 10 random 1000 bp sequences with the same di-nucleotide frequencies as those of the 1000 bp around the SNP using the fasta-shuffle-letters program from MEME Suite[34] to serve as a null background set. For each null sequence, we created a null effect allele sequence and a null non-effect allele sequence by replacing the base at the center of the null sequence with the effect and non-effect allele, respectively.

For each SNP in a peak in each of the clusters, we computed GkmExplain importance scores for each of the central 200 bp in each of the 10 null effect and non-effect allele sequences using each of the 10 gkm-SVM models for the respective clusters. Next, for each pair of null sequences, we subtracted the sum of the importance scores corresponding to the central 50 bp of the null non-effect allele sequence from that of the null effect allele sequence to compute the null GkmExplain score.

To compute null in silico mutagenesis (ISM) scores for each SNP in a peak in each of the clusters, we used each of the 10 fold gkm-SVM models from the respective clusters to compute model output prediction scores for the central 50 bp of the null effect and non-effect allele sequences by running gkmpredict. Then, we subtracted the score of the null non-effect allele sequence from the score of the null effect allele sequence to obtain the null ISM score.

To compute null deltaSVM scores, for each SNP in a peak in each of the clusters, we used each of the 10 sets of *k*-mer scores from the 10 fold gkm-SVM models from the respective cluster to run deltaSVM on the central 50 bp of the null effect and non-effect allele sequences.

We found that the t-distribution was a good fit (based on KS test) to the empirical null distribution for all three scores. Hence, we used the fitted t-distributions (using SciPy python library http://www.scipy.org/) to each of the three sets of null scores as the null distributions.

To select SNPs with **statistically significant gkm-SVM allelic scores**, for each cluster, we selected those SNPs that fall outside the 95% confidence interval for all three null *t*-distributions fitted to the GkmExplain, ISM, and deltaSVM scores.

Next, we developed a method to identify putative transcription factor binding sites around each gkm-SVM scored statistically significant candidate SNP, by identifying the subsequences around the SNP whose base-resolution importance scores are significantly above that of the di-nucleotide matched shuffled background. We use the GkmExplain importance scores of all bases in the central 200 bp of all the null effect and non-effect allele sequences as a null distribution to identify bases around the SNP with high signal-to-noise ratio. For each SNP, we defined the **active allele** as the allele for which the 50 bp sequence centered on the SNP has the higher sum of non-negative importance scores relative to the other allele. Next, starting from the center of the active allele's sequence, which is the location of the SNP, we continue advancing one pointer upstream and another downstream, each up to the position beyond which lie two consecutive bases that both have consensus importance scores that are not higher than 97.5% of the null importance scores. The subsequence between the terminal positions of the two pointers corresponds to one that underlies a series of bases with high GkmExplain importance scores that are significantly above the null scores of the di-nucleotide matched shuffled background sequences and potentially contains transcription factor binding sites and motifs that are relevant for the given cluster. We refer to these high-importance subsequences as seqlets. If a SNP does not have a seqlet that reaches a minimum length of 7 bp, then we alternatingly extend each end of the seqlet by 1 bp until this minimum length is reached.

Next, we defined two additional scores (prominence score and magnitude score) to further identify high confidence candidates from the gkm-SVM scored statistically significant candidate SNPs that are supported by seqlets that could potentially match identifiable transcription factor binding sites. We compute the sum of the non-negative consensus importance scores from the positions of the effect allele that overlap the active allele's seqlet, which we refer to as the **effect allele seqlet score**, and divide that score by the sum of the non-negative consensus importance scores from the entire central 200-bp region of the effect allele sequence; we refer to this ratio as the **effect allele seqlet signal-to-noise ratio**. Similarly, we compute the **non-effect allele seqlet score** as the sum of the non-negative consensus importance scores in the non-effect allele sequence from the same positions overlapping the active seqlet. We obtain a corresponding **non-effect allele seqlet signal-to-noise ratio** by dividing the non-effect allele seqlet score by the sum of the non-negative consensus importance scores from the entire central 200-bp region of the non-effect allele sequence. Then, for each SNP, we compute the **prominence score** by subtracting the non-effect allele seqlet signal-to-noise ratio from the effect allele seqlet signal-to-noise ratio. In addition, we also compute a **magnitude score** by subtracting the non-effect allele seqlet score from the effect allele seqlet score.

To compute the statistical significance of the prominence and magnitude scores for candidate SNPs, for each cluster, we compute null prominence scores and null magnitude scores for each pair of null effect and non-effect allele sequences using the same procedure described above and use the empirical null distributions to obtain *p*-values for the prominence and magnitude scores for each candidate SNP scored for that cluster. For each type of score, in order to control for any arbitrary bias in the sign of the score, we include the negative value of each score to the list of scores to enforce symmetry before fitting the distribution.

Finally, to prioritize SNPs that disrupt potential transcription factor binding sites, in each cluster, among the SNPs with statistically significant gkm-SVM allelic scores, we designate as high confidence SNPs those that have a prominence score with a *p*-value less than 0.05. These are the SNPs that have an allele that completely destroys a prominent and high-scoring seqlet and, as a result, potentially disrupts an important transcription factor binding site. Next, among the confident SNPs that do not pass the high confidence threshold, we designated as medium confidence SNPs those that have either a magnitude score with a *p*-value less than 0.05 or a prominence score with a *p*-value less than 0.10. The magnitude score threshold is intended to capture those SNPs that have a significant deleterious effect on the seqlet score, even if those SNPs do not necessarily destroy the entire seqlet and even for cases where the seqlet around the SNP is not among the most prominent seqlets in the local 200 bp sequence window. In addition, the relaxed prominence score threshold is intended to capture those SNPs that do not pass the stringent filter for the high confidence set, but nevertheless, demonstrate at least a partial deleterious effect on a moderately scoring seqlet around the SNP. Together, these two filters serve to increase the recall in the prioritization of the SNPs, allowing us to identify all promising SNPs that are worthy of in-depth evaluation, which can assess their potential regulatory effect through a case-by-case analysis. The remaining SNPs in the confident set, which fail to meet the threshold for medium confidence, are designated as low confidence SNPs, as they include SNPs that significantly reduce the GkmExplain score, the ISM score, and the deltaSVM score, but do not have a clear impact on a seqlet around the SNP, making it unlikely for them to have a disruptive effect on a key transcription factor binding site.

SUPPLEMENTARY INFORMATION REFERENCES

1. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).
2. Hirata, T. *et al.* Zinc-finger genes Fez and Fez-like function in the establishment of diencephalon subdivisions. *Development* **133**, 3993–4004 (2006).
3. Yuan, L. L. & Chen, X. Diversity of potassium channels in neuronal dendrites. *Prog. Neurobiol.* **78**, 374–389 (2006).
4. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 1–10 (2015).
5. Corces, M. R. *et al.* Lineage-specific and single cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
6. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
7. van Bruggen, D., Agirre, E. & Castelo-Branco, G. Single-cell transcriptomic analysis of oligodendrocyte lineage cells. *Curr. Opin. Neurobiol.* **47**, 168–175 (2017).
8. Molofsky, A. V. *et al.* Astrocyte-encoded positional cues maintain sensorimotor circuit integrity. *Nature* **509**, 189–194 (2014).
9. Zhao, T. *et al.* Genetic mapping of Foxb1-cell lineage shows migration from caudal diencephalon to telencephalon and lateral hypothalamus. *Eur. J. Neurosci.* **28**, 1941–1955 (2008).
10. Bosse, A. *et al.* Identification of the vertebrate Iroquois homeobox gene family with overlapping expression during early development of the nervous system. *Mech. Dev.* **69**, 169–181 (1997).
11. Fong, W. L., Kuo, H. Y., Wu, H. L., Chen, S. Y. & Liu, F. C. Differential and Overlapping Pattern of Foxp1 and Foxp2 Expression in the Striatum of Adult Mouse Brain. *Neuroscience* **388**, 214–223 (2018).
12. 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. 2019 513* **51**, 414 (2019).
13. Deng, H. X. *et al.* Identification of TMEM230 mutations in familial Parkinson's disease. *Nat. Genet.* **48**, 733–739 (2016).
14. Nott, A. *et al.* Brain cell type – specific enhancer – promoter interactome maps and disease-risk association. *Science* **1139**, 1134–1139 (2019).
15. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics reveals novel disease risk mechanisms. *biorxiv* (2019) doi:10.1101/694281.
16. Hyman, B. T. *et al.* National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimer's Dement.* **8**, 1–13 (2012).
17. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).
18. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
19. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nat. Genet.* **37**, 129–137 (2005).
20. 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).
21. Lai, M. C. *et al.* Haplotype-specific MAPT exon 3 expression regulated by common intronic polymorphisms associated with Parkinsonian disorders. *Mol. Neurodegener.* **12**, 1–16 (2017).
22. Allen, M. *et al.* Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimer's Res. Ther.* **6**, 1–14 (2014).
23. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
24. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
25. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

26. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

27. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).

28. Liu, B. *et al.* Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. *Am. J. Hum. Genet.* **103**, 377–388 (2018).

29. Lee, D. LS-GKM: A new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).

30. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

31. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: Fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).

32. Ghandi, M. *et al.* GkmSVM: An R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).

33. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).

34. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202–208 (2009).