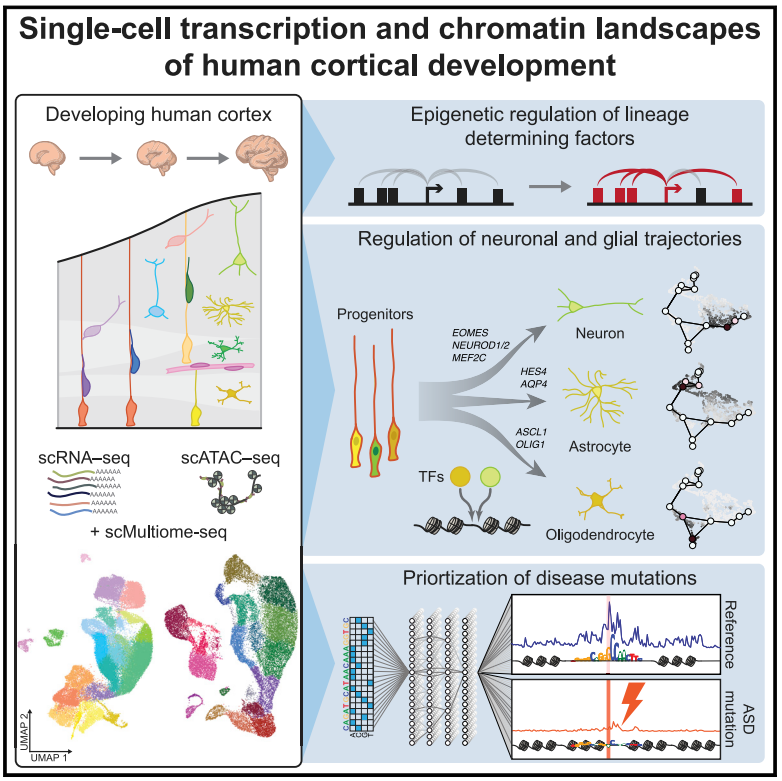
Cell

Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution

Graphical abstract



HIGHLIGHTS

- Single-cell RNA and chromatin profiling charts human corticogenesis
- Distinct TFs underlie neurogenesis and gliogenesis regulatory programs
- Lineage-determining TFs adopt an active chromatin state early in differentiation
- Neural networks prioritize noncoding *de novo* mutations in autism spectrum disorder

Authors

Alexandro E. Trevino, Fabian Müller, Jimena Andersen, ..., Anshul Kundaje, Sergiu P. Paşca, William J. Greenleaf

Correspondence

spasca@stanford.edu (S.P.P.), wjg@stanford.edu (W.J.G.)

In brief

A single-cell atlas of gene expression and chromatin accessibility of human developing cortex during mid-gestation reveals lineage-determining transcription factors for human corticogenesis and identifies prioritized mutations for autism spectrum disorder.

Trevino et al., 2021, Cell *184*, 5053–5069
September 16, 2021
https://doi.org/10.1016/j.cell.2021.07.039

CellPress

Cell

CellPress

Resource

Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution

Alexandro E. Trevino,[1,13] Fabian Müller,[1,2,13] Jimena Andersen,[3,4,13] Laksshman Sundaram,[5,13] Arwa Kathiria,[1] Anna Shcherbina,[6] Kyle Farh,[7] Howard Y. Chang,[1,8,9] Anca M. Paşca,[10] Anshul Kundaje,[1,5] Sergiu P. Paşca,[3,4,14,*] and William J. Greenleaf[1,11,12,*]

[1]Department of Genetics, Stanford University, Stanford, CA, USA
[2]Center for Bioinformatics, Saarland University, Saarbrücken, Germany
[3]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
[4]Stanford Brain Organogenesis Program, Wu Tsai Neuroscience Institute Stanford University, Stanford, CA, USA
[5]Department of Computer Science, Stanford University, Stanford, CA, USA
[6]Biomedical Data Science Program, Stanford University, Stanford CA, USA
[7]Illumina Artificial Intelligence Laboratory, Illumina Inc, San Diego, CA, USA
[8]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, USA
[9]Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA
[10]Department of Pediatrics, Division of Neonatology, Stanford University, Stanford, CA, USA
[11]Department of Applied Physics, Stanford University, Stanford, CA, USA
[12]Chan–Zuckerberg Biohub, San Francisco, CA, USA
[13]These authors contributed equally
[14]Lead contact
*Correspondence: spasca@stanford.edu (S.P.P.), wjg@stanford.edu (W.J.G.)
https://doi.org/10.1016/j.cell.2021.07.039

SUMMARY

Genetic perturbations of cortical development can lead to neurodevelopmental disease, including autism spectrum disorder (ASD). To identify genomic regions crucial to corticogenesis, we mapped the activity of gene-regulatory elements generating a single-cell atlas of gene expression and chromatin accessibility both independently and jointly. This revealed waves of gene regulation by key transcription factors (TFs) across a nearly continuous differentiation trajectory, distinguished the expression programs of glial lineages, and identified lineage-determining TFs that exhibited strong correlation between linked gene-regulatory elements and expression levels. These highly connected genes adopted an active chromatin state in early differentiating cells, consistent with lineage commitment. Base-pair-resolution neural network models identified strong cell-type-specific enrichment of noncoding mutations predicted to be disruptive in a cohort of ASD individuals and identified frequently disrupted TF binding sites. This approach illustrates how cell-type-specific mapping can provide insights into the programs governing human development and disease.

INTRODUCTION

Dynamic changes in activity of *cis*-regulatory DNA elements, driven by changes in transcription factor (TF) binding, underlie phenotypic transformations during development (Buenrostro et al., 2018; Stergachis et al., 2013). Single-cell methods for measuring chromatin accessibility have emerged as a sensitive probe for this activity and, combined with tools to measure single-cell transcriptomes, have the potential to decipher how combinations of TFs drive gene expression programs (Kelsey et al., 2017; Klemm et al., 2019). Quantifying the dynamic activity of regulatory elements also enables the inference of the time point or cell type wherein disease-associated genetic variation impacts development. For instance, it is still unknown how genetic variants associated with autism spectrum disorder (ASD) interfere with the genetic programs underlying the development of the cerebral cortex (Rubenstein, 2011; Zhou et al., 2019).

Corticogenesis is a dynamic, highly regulated process characterized by the expansion of apical and basal radial glia (RG) and intermediate progenitors in the ventricular and subventricular zones (VZ, SVZ), the inside-out generation of glutamatergic neurons, and the differentiation of astrocytes and oligodendrocytes (Greig et al., 2013; Molnár et al., 2019; Silbereis et al., 2016). Cell types derived outside of the dorsal forebrain, including GABAergic neurons, microglia, and some oligodendrocytes, also migrate and integrate into the cortex (Wonders and

Anderson, 2006). Resolving gene-regulatory dynamics associated with these developmental trajectories requires investigation of both chromatin and gene expression states at single-cell resolution.

To map the gene-regulatory logic of human corticogenesis, we generated single-cell chromatin accessibility and RNA expression profiles from human fetal cortical samples spanning 8 weeks during mid-gestation. These paired maps revealed a class of genes with comparatively large numbers of nearby putative enhancers whose accessibility was strongly predictive of gene expression. These genes with predictive chromatin (GPCs) are frequently TFs, and we observed that their local accessibility precedes lineage-specific gene expression in cycling progenitors. We validated these findings using single-cell accessibility and expression profiles derived from the same cell (multiomics). We defined a developmental trajectory for cortical glutamatergic neurons, revealing a continuous progression of TF motif activities associated with neuronal specification and migration, and explored the co-dependencies in TF motif accessibility along this trajectory. In addition, we characterized the lineage potential of glial progenitors and provided evidence for two distinct astrocyte precursor subtypes. Finally, we trained a deep-learning model to infer base-pair-resolved, cell-type-specific chromatin accessibility profiles from DNA sequence. These models allowed prediction of the potential impact of genetic variants on the cell-type-specific chromatin landscape and prioritized rare *de novo* genetic variants associated with ASD, demonstrating the ability to map disease risk with single-cell and single-base resolution during cortical development.

RESULTS

A single-cell regulatory atlas of the developing human cerebral cortex

To capture cellular heterogeneity in the cerebral cortex, we created a gene-regulatory atlas using the Chromium platform (10x Genomics) to generate single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq) and single-cell RNA sequencing (scRNA-seq) libraries from four primary samples at post-conceptional week (PCW) 16, PCW20, PCW21, and PCW24 (Figure 1A). Overall, we obtained 57,868 single-cell transcriptomes and 31,304 single-cell epigenomes after quality control and filtering (Table S1; Figures S1A–S1H). Consistent with previous studies (Fietz et al., 2010; Hansen et al., 2010; Kang et al., 2011; Pollen et al., 2015; Trevino et al., 2020), $CTIP2^+$ cells were present in the cortical plate (CP) and $SOX9^+$ cells in the VZ, SVZ, and outer SVZ (oSVZ; Figures 1B and S1I), while the $GFAP^+$ scaffolding spanned the neocortex at PCW17 and PCW21 (Figures 1C and S1J). The proliferation marker KI67 colocalized with both $GFAP^+$ cells and with $PPP1R17^+$ intermediate progenitor cells (IPCs) in the SVZ and oSVZ (Figures 1C and S1J).

To assess global similarities and differences between individual cells, we performed unsupervised analyses, including dimension reduction using uniform manifold approximation and projection (UMAP) and clustering. For scATAC-seq, we employed an iterative approach (Granja et al., 2019) to obtain a low-dimensional embedding, cell clustering, and a consensus set of 657,930 accessible peaks representing potential *cis*-regulatory elements (CREs; STAR Methods). The structure of the RNA and chromatin representations were similar, with variation related to gestational time (Figure 1D) and cell types. Performing both assays on the same samples enabled us to dissect complementary aspects of gene regulation, including the relationship between gene expression (scRNA-seq) and chromatin accessibility-based gene activity score (scATAC-seq)—a metric defined by the aggregate local chromatin accessibility of genes (hereafter "gene activity score"; STAR Methods) (Pliner et al., 2018) as well as aggregate TF motif activity scores (Schep et al., 2017). Corticogenesis TFs such as *SOX9*, *EOMES*, *NEUROD2*, and *DLX2* showed strong cluster-specific enrichments in these three metrics (Figure 1E) consistent with their ascribed roles in RG, IPCs, cortical glutamatergic neurons (GluN), and GABAergic neurons (interneuron; IN), respectively.

We next called clusters in both datasets (Figure 1F; STAR Methods) and annotated these clusters using gene expression and gene activities of known markers (Lui et al., 2011; McConnell, 1995; Nowakowski et al., 2016, 2017; Polioudakis et al., 2019; Pollen et al., 2015; Thomsen et al., 2016) (Figures 1G–H, S2A, and S2B; Table S1; STAR Methods). In scRNA-seq, we observed a cluster of cycling cells (Cyc) expressing *TOP2A* and *KI67*. We also found that RG, expressing *SOX9* and *HES1*, included both ventricular radial glia (vRG: *FBXO32*, *CTGF*) and outer radial glia (oRG: *MOXD1*, *HOPX*), and these were separated according to time (early RG, PCW16: *NPY*, *FGFR3*; late RG, PCW20–24: *CD9*, *GPX3*). Cells in one scRNA-seq cluster expressed markers for truncated RG (tRG) and ependymal cells (tRG: *CRYAB*, *NR4A1*, *FOXJ1*). We also identified a cluster expressing genes associated with both RGs and oligodendrocyte lineage precursors (*ASCL1*, *OLIG2*, *PDGFRA*, *EGFR*). This cluster, which we named multipotent glial progenitor cells (mGPC), was different from the OPC and oligodendrocyte (OPC/Oligo) cluster that expressed *SOX10*, *NKX2.2*, and *MBP*. Genes associated with astrocyte identity (*AQP4*, *APOE*) were observed in the mGPC cluster as well as in the late RG cluster. A large domain was composed of neuronal IPC (*EOMES*, *PPP1R17*, *NEUROG1*) and GluN (*BCL11B*/*CTIP2*, *SATB2*, and *SLC17A7*/*VGLUT1*). Among the GluN clusters, we found cells expressing subplate markers (SP: *NR4A2*, *CRYM*). We also identified distinct clusters of IN expressing *DLX2* and *GAD2*—one of them expressed markers associated with medial ganglionic eminence (MGE: *LHX6*, *SST*) and the other expressed markers associated with both caudal ganglionic eminence and pallial-subpallial boundary (CGE: *SP8*, *NR2F2*; PSB: *MEIS2*, *ETV1*). In addition, we observed clusters of microglia (MG: *AIF1*, *CCL3*), endothelial cells (EC: *CLDN5*, *PECAM1*), pericytes (Peric: *FOXC2*, *PDGFRB*), leptomeningeal cells (VLMC: *COL1A1*, *LUM*), and red blood cells (RBC: *HEMGN*). Many of the above markers exhibited dynamic gene activity scores in corresponding clusters in scATAC-seq space (Figure 1H). While most clusters had cells representing all time points, some were strongly biased for earlier or later stages (e.g., mGPCs and tRGs; Figure S2C). To further corroborate cell-type identities and gestational time, we projected two previously published scRNA-seq datasets from human cortex into our scRNA-seq manifold (Bhaduri et al., 2020; Polioudakis et al., 2019). We computed Jaccard

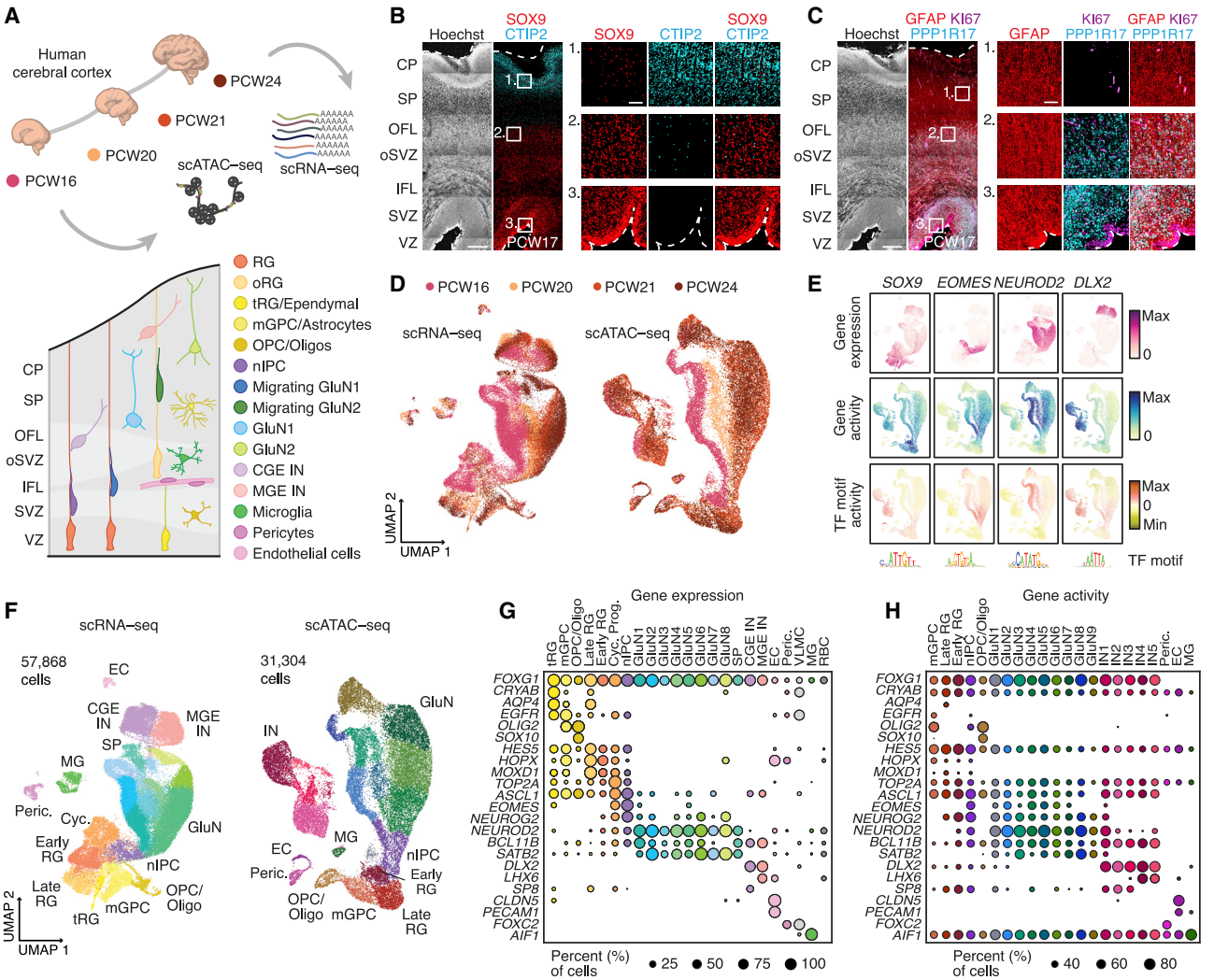


Figure 1. A single-cell epigenomic atlas of the human cerebral cortex

(A) Schematic of time, profiling methods, and cell types.

(B) Immunohistochemistry for SOX9 and CTIP2 at PCW17. VZ, ventricular zone; SVZ, subventricular zone; IFL, inner fiber layer; oSVZ, outer SVZ; OFL, outer fiber layer; SP, subplate; CP, cortical plate. This image was generated by automatic stitching of individual images.

(C) Immunohistochemistry for GFAP, KI67, and PPP1R17 at PCW17. This image was generated by automatic stitching of individual images.

(D) UMAP based on gene expression (left) and peak accessibility (right). Cells colored according to time.

(E) Multimodal profiling of *SOX9*, *EOMES*, *NEUROD2*, and *DLX2* including gene expression (scRNA-seq), gene activity scores, and TF motif activity (scATAC-seq).

(F) UMAP of cells colored by cluster. RG, radial glia; Cyc, cycling progenitors; tRG, truncated radial glia, mGPC, multipotent glial progenitor cell; OPC/Oligo, oligodendrocyte progenitor cell/oligodendrocyte; nIPC, neuronal intermediate progenitor cell; GluN, glutamatergic neuron; CGE IN, caudal ganglionic eminence interneuron; MGE IN, medial ganglionic eminence interneuron; EC, endothelial cell; MG, microglia; Peric., Pericytes

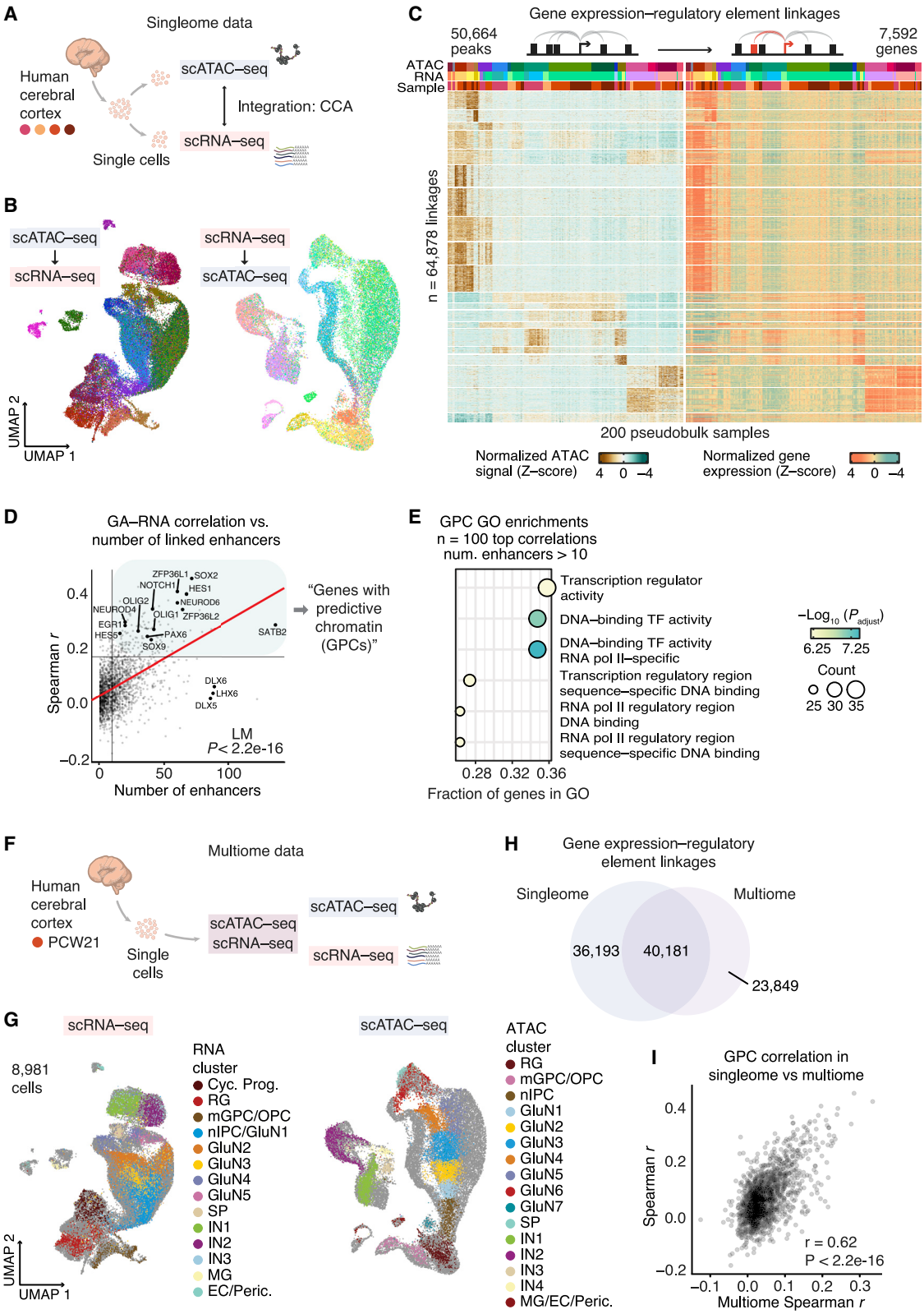
(G) Dotplot showing the cells expressing selected markers across scRNA-seq clusters.

(H) Dotplot showing marker gene activity scores derived across scATAC-seq clusters.

Scale bars, 500 μm (B, C), 100 μm (insets B, C).

indices of correspondence and observed high agreement between cell types, cell-cycle phase, and gestational times in our data and the computationally matched independent annotation (Figures S3A–S3G).

We integrated the derived gene activity scores with gene expression levels using canonical correlation analysis (CCA) to match cell data from each modality to the nearest neighbors in the other data representation (Figure 2A) (Stuart et al., 2019). Cluster annotations of matched cells were consistent, except for the cycling progenitor cluster in scRNA-seq, which did not directly map to cells in the chromatin landscape (Figures 2B, S3H, and S3I). Using pseudo-bulk aggregates of these matched annotations, we applied a correlation-based approach that links gene-distal CRE accessibility to gene expression (Corces et al.,



(legend on next page)

2018; Ma et al., 2020; Trevino et al., 2020), identifying 64,878 CRE-gene pairs that represent potential enhancer-gene interactions (Table S2). In this analysis, a gene was linked to a median of five CREs, and linked CREs were more conserved than unlinked elements (Figure S3J, Wilcoxon rank-sum test $p < 2.2 \times 10^{-16}$) and more likely to be supported by cell-type-specific three-dimensional (3D) interactions from a recently published promoter-centric chromosome conformation capture dataset (Song et al., 2020) (Figures S3K–S3M). Co-variation of CRE accessibility and gene expression distinguished the identified cell types in both scRNA-seq and scATAC-seq (Figure 2C). Clustering the associated CRE accessibility revealed particularly high variability across clusters corresponding to glial cell populations, corroborated the distinctiveness of IN clusters, and indicated dynamic patterns of gene regulation across GluN clusters.

We then identified genes whose expression could be well predicted from local single chromatin accessibility by ranking gene activity-expression correlations. Genes with the highest correlation included *SOX2* and *HES1*, and these genes were linked to greater numbers of putative enhancers. We hypothesized that these comprised a class of highly regulated genes that play a driving role in establishing cell identities in the developing cortex and defined a set of 185 genes with predictive chromatin (GPCs; genes in the top decile of gene activity-expression correlations, linked to >10 CREs) (Table S2; Figure 2D). This gene set was strongly enriched for transcription regulator activity and DNA-binding TF activity (Figure 2E).

To validate these inferences, we profiled scATAC-seq and scRNA-seq data from the same cells in PCW21 human cortex (multiome) (Figure 2F). Filtering across both data modalities resulted in 8,981 cells with high-quality transcriptome and epigenome profiles (Table S2; Figures S3N–S3T). We projected multiomic scATAC-seq and scRNA-seq profiles into the corresponding individually generated landscapes and confirmed that our cell-type annotations were well represented in the joint data (Figure 2G). Applying our CRE-gene linking approach to the true cell-to-cell matches, we found that 40,181 inferred peak-gene linkages (53%) were observed from this single time point measurement and an additional 23,849 were identified (Figure 2H; Table S2), demonstrating that most inferred CRE-gene interactions were observed in this joint dataset. Similarly, we applied CCA to multiome data, where the correct cell assignments are known. The inferences were generally validated by the true clusters, and this agreement was increased by assigning clusters based on 50 nearest neighbors in CCA space, rather than the single closest neighbor (Figures S4A and S4B). In addition, we found a strong concordance in GPC activity-expression correlations of *in silico*-linked singleome cells versus multiome cells (Figure 2I). GPCs are thus also readily apparent in this joint dataset, underlining the correspondence between their local accessibility and their transcription within the same cell.

Continuous trajectories of gene regulation across cortical neuron differentiation

GluN are born in a specific sequence during development. Although several key factors controlling their fate have been described (Greig et al., 2013), the logic that governs their specification, migration, and maturation has not been resolved in human development. To define a trajectory of GluN development, we annotated each cell in associated clusters with pseudotime values. This annotation was derived using an algorithm based on diffusion through cell-similarity networks derived from RNA velocity (Bergen et al., 2020; La Manno et al., 2018) (Figures 3A and S4C–S4F). Notably, the algorithm rooted the trajectory in the cluster of cycling cells (Figure S4D). To test how the architecture of the cortex mapped onto this trajectory, we next projected an independent scRNA-seq data comprising adult cortical neurons (Hodge et al., 2019) into the developmental landscape and identified the nearest neighbor cell for each adult scRNA-seq profile (Figure S4G). Adult GluN projected preferentially into the neighborhoods of cells annotated with later pseudotimes, and pseudotime was significantly associated with the annotated layer of adult cells (one-sided Wilcoxon rank-sum test $p < 9.6 \times 10^{-15}$; Figure S4H). As expected, we also observed significant association of earlier and later time points with deep and superficial adult cortical layers (one-sided Fisher's exact test $p < 2.7 \times 10^{-124}$; Figure S4I). When we compared the expression levels in migrating neurons from the early gestational time point (PCW16) to those from later time points (PCW20–PCW24), we observed increased expression of *LIMCH1*, *RUNX1*, *SNCB*, and *DOK5* and decreased expression of the AP-1 TF family (*JUN*, *FOS*), heat shock factors *HSPA1A/B* and *DUSP1* (Figures S4J and S4K; Table S3). Overall, we found surprisingly few differentially expressed genes previously implicated in neurogenesis, suggesting a considerable degree of gene expression and regulatory variability could be associated with pseudotime rather

Figure 2. Integrative and multiomic gene regulatory dynamics in the human cortex

(A) Generation and integration of singleome scATAC-seq and scRNA-seq data. CCA, canonical correlation analysis.

(B) UMAPs of scRNA-seq and scATAC-seq cells colored by cluster assignment of matched cells.

(C) Heatmap showing chromatin accessibility and gene expression of 64,878 significantly linked CRE-gene pairs. Shown are side-by-side heatmaps in which one row represents a pair of one CRE and one linked gene). Each CRE can be linked to multiple genes, and each gene can be linked to multiple CREs. Hence, each gene and each CRE can be represented by multiple rows in the corresponding heatmap. Pairs (rows) were clustered using k-means clustering (k = 20). For visualization, 10,000 rows were randomly sampled. Columns represent 200 pseudobulk samples, which have been annotated using the majority RNA cluster, ATAC cluster, and time of all cells in the pseudobulk.

(D) Correlation between single-cell gene expression and chromatin-derived gene activity scores (GA), and the number of linked CREs per gene. TFs are labeled.

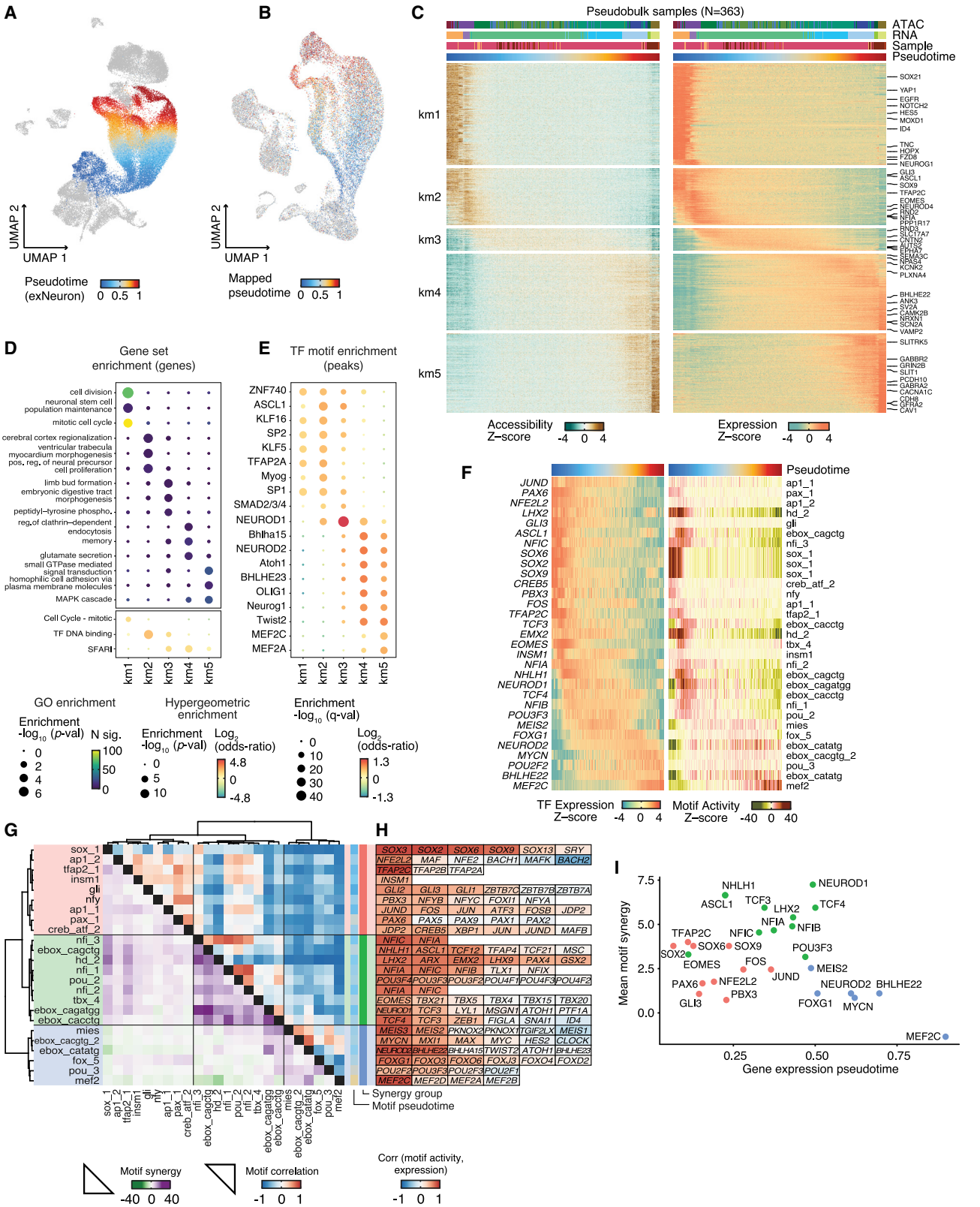
(E) GO enrichment analysis of the 185 genes with GPCs in (D).

(F) Generation of scATAC-seq and scRNA-seq data from the same cells (multiome data).

(G) Projection of multiome scATAC-seq into singleome scATAC-seq UMAP space, and multiome scRNA-seq into singleome scRNA-seq UMAP space. Cell coloring corresponds to multiome cluster assignment, and gray cells show the singleome manifold onto which multiome cells have been projected.

(H) Venn diagram showing overlap of CRE-gene linkages identified in singleome versus multiome data.

(I) Correlation showing the correspondence between predictive chromatin in singleome versus multiome data.



(legend on next page)

than gestational time. We therefore decided to investigate the regulatory dynamics along the pseudotime axis.

To connect expression trajectories to the accessibility dynamics of regulatory elements, we transferred pseudotime values from RNA cells to their nearest ATAC cell neighbors, confirming this produced a smooth continuum of pseudotime in the chromatin manifold (Figure 3B). By applying our correlation-based CRE-to-gene linking approach to the glutamatergic neuronal lineage, we identified 13,989 dynamic interactions across pseudotime and grouped these into five clusters (Figure 3C; Table S3). Linked genes active early in pseudotime exhibited gene ontology (GO) enrichments for cell division and neural precursor proliferation, whereas later interactions were associated with morphogenesis, migration, and maturation (Figure 3D). Interestingly, genes encoding TFs and DNA-binding proteins were particularly enriched in intermediate interactions, while genes implicated in ASD susceptibility (Abrahams et al., 2013) were more likely to be linked later in pseudotime. To nominate TFs that may control these programs, we identified motifs that were enriched in the different clusters of linked regulatory elements. Motifs enriched in interactions early in the trajectory included ZNF740, KLF16, SP1/2, and ASCL1 (Figure 3E). Conversely, interaction clusters associated with intermediate and late pseudotime were associated with motifs of neuronal TFs (NEUROD1/2, NEUROG1, MEF2C).

We next characterized the TF-driven regulatory dynamics of neurogenesis over pseudotime. To mitigate correlation biases due to sequence similarity between motifs in this analysis, we utilized a resource of previously disambiguated clusters of TF motifs (Vierstra et al., 2020). We then linked specific TF genes to these motif clusters by correlating TF expression with the accessibility-derived motif activity scores, resulting in pairings of 31 TFs and 24 motif clusters (STAR Methods). We observed synchronized TF expression and motif activity for dynamic regulators along developmental pseudotime, starting with PAX6, SOX2/6/9, GLI3, and ASCL1 motifs, followed by intermediate stage factor motifs (EOMES, NFIA, NFIB, NEUROD1), and finally late-stage motifs (NEUROD2, BHLHE22, MEF2C; Figure 3F). Together, these data describe cohesive, sequential waves of motif activation during corticogenesis that are consistent across gestational time points.

To understand how TFs are coordinated during corticogenesis, we computed the genome-wide synergy and correlation patterns of motif family accessibility (Figures 3G and 3H; STAR Methods) (Schep et al., 2017). We found three broad classes of motifs that associated with accessibility and TF expression over pseudotime (Figures 3G–3I): (1) early-activity motifs exhibiting moderate synergies (SOX, GLI, PAX), (2) intermediate activity motifs (NFI, TBX/EOMES) that are highly synergetic within their class, and (3) late-activity motifs that are less cooperative and generally appear to operate more independently (NEUROD2/BHLHE22, MEF2). These findings suggest a higher degree of TF motif coordination early in neurogenesis and regulation of maturation by a smaller set of more independent TFs.

Clustering approach to link gene expression programs to cell-fate decisions

We observed extensive heterogeneity in glial populations, corresponding to distinct yet partially overlapping expression programs in the identified clusters (Figures S5A and S5B). We adopted an analysis to identify modules of co-expressed genes using fuzzy c-means clustering (STAR Methods; Figures 4A and S5C; Table S4), allowing cells to be annotated with module activities, and for genes to be shared between multiple modules (Figures S5C and S5D; Table S4), enabling analysis of how cells may progress from one module to another. We projected these cell loadings into a low-dimensional representation of differentiation (Figure 4A, bottom). The structure of this embedding and the underlying module assignments was stable to fuzzy clustering parameters (STAR Methods).

To understand the biological basis of these gene modules, we examined their expression across cell clusters, developmental stage, and pseudotime (Figure 4B), which was rooted in cycling (“Cyc”) cells and correlated with time (Figure S5E; STAR Methods). Glial maturation genes *FOXJ1*, *AQP4*, and *MBP*, which are markers for ciliated ependymal cells, astroglia, and oligodendrocytes, respectively (Barbarese et al., 1988; Jacquet et al., 2009; Zhang et al., 2016), were expressed in late-pseudotime cells and assigned to modules m5, m2, and m7. In contrast, the expression of genes associated with cell division and progenitors, such as *TOP2A*, *NR2F1*, and *NFIC*, peaked early in pseudotime and were assigned to modules m10, m6, and m3 (Figures 4C and S5F). Some modules (m6, m8) spanned many samples and stages, indicative of sustained expression programs, while others were restricted (m5, m14; Figures 4B, 4D, and S5G). Modules exhibited distinct GO enrichments, including “cation and metal ion binding” in m6, which may be related to the role of human astrocytes in metal homeostasis (Vasile et al., 2017; Zhang et al., 2016) and disease (Figures S5H and S5I). Module m5, comprising *FOXJ1*$^{+}$ cells, was enriched for dynein

Figure 3. Molecular signatures of cortical glutamatergic neurons

(A) UMAP of scRNA-seq cells highlighting the GluN trajectory and RNA-velocity derived pseudotime.

(B) UMAP of scATAC-seq with transferred pseudotime annotation.

(C) Heatmap showing chromatin accessibility and gene expression of 13,989 significantly linked CRE-gene pairs (columns: left, CRE accessibility; right, linked gene expression) across 363 pseudobulk samples aggregated along pseudotime bins in the GluN trajectory. Interactions (rows) were clustered using k-means clustering (k = 5).

(D) Gene set enrichment analysis of genes represented in the five interaction clusters.

(E) Enrichment of TF motifs in peaks represented in the five interaction clusters. Color represents odds ratios; size represents the $-\log_{10}$ (p value).

(F) Heatmaps showing *Z* score normalized expression (left) and motif activity (right) of TFs in 363 pseudobulk samples aggregated along pseudotime bins. Rows show 31 dynamic TFs associated with 24 motif clusters.

(G) TF motif correlation coefficients (upper) and synergy *Z* scores (lower) of the 24 motif clusters in (F).

(H) Correlation coefficients of TF motif cluster chromatin activity and annotated gene expression.

(I) Scatterplot showing aggregate gene expression pseudotime versus mean motif synergy. Point colors denote the cluster assignments in (G).

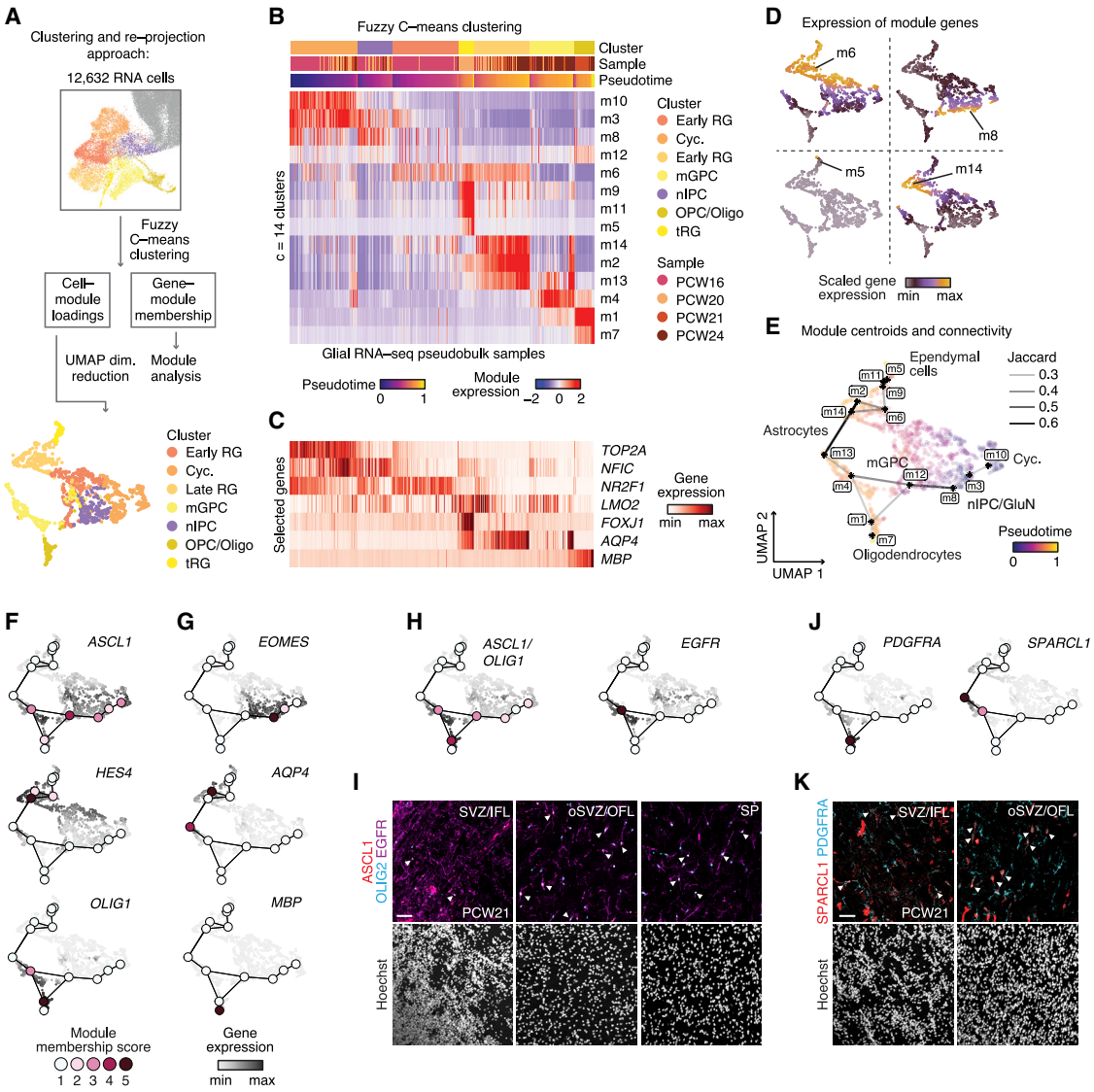


Figure 4. Regulatory logic of glial cell specification

(A) Schematic illustrating the expression-based clustering and reprojection of glial cells. Points in the bottom panel correspond to pseudobulk aggregates of 50 cells.

(B) Heatmap of module expression across pseudobulk aggregates, showing variation by cluster, sample age, and pseudotime.

(C) Heatmap showing the expression of selected genes across the same pseudobulks.

(D) Mean scaled expression in the low-dimensional UMAP embedding of selected gene modules. Figure S5G shows all modules.

(E) Projection of module centroids into UMAP space. Pseudobulk samples are colored by pseudotime. Module overlap is shown by links between centroids and was computed by thresholding the pairwise Jaccard index at >0.2.

(F) Module membership and expression values for *ASCL1*, *HES4*, and *OLIG1*.

(G) Module membership and expression values for *EOMES*, *AQP4*, and *MBP*.

(H) Module membership and expression values for *ASCL1/OLIG1* and *EGFR*.

(I) Immunohistochemistry showing expression and colocalization (white arrowheads) of ASCL1, OLIG2, and EGFR in cells of the SVZ, oSVZ, outer and inner fiber layers (OFL, IFL), and SP.

(J) Module membership and expression values for the oligodendrocyte progenitor marker *PDGFRA* and the astrocyte-associated gene *SPARCL1*.

(K) Immunohistochemistry showing expression and colocalization (white arrowheads) of SPARCL1 and PDGFRA in cells of the SVZ, oSVZ, and outer and inner fiber layers (OFL, IFL).

Scale bars, 50 μm (J and K).

binding and microtubule activity, consistent with the role in circulating cerebrospinal fluid (Ransom, 2012). Immunohistochemistry revealed that TFAP2C, which associated with module m6, was expressed in VZ and SVZ (Figures S6A and S6B). Similarly, PBXIP1, which was associated with m2, was expressed in RG in the VZ and SVZ but not in more mature CP astrocytes (Figures S6C and S6D). CRYAB, associated with m9, was expressed in tRG in the VZ, as described (Figures S6E and S6F) (Nowakowski et al., 2016).

Our clustering and reprojection approach enabled us to compute the degree of gene overlap between modules, which provided a measure of module similarity across our glial landscape (Figure S6G). To visualize these relationships, we computed the weighted average of module gene expression across pseudobulk aggregates and plotted these ''module centroids'' and their connectivity in the embedding, along with pseudobulks and their pseudotime values (Figure 4E). Investigation of module memberships in this representation revealed three broad programs emanating from the cycling cluster: (1) an *ASCL1*$^+$ program associated with m3 and m8 and terminating in *EOMES*$^+$ nIPCs; (2) a *HES4*$^+$ program associated with module m6 and terminating in astrocytes and ependymal cells; and (3) an *ASCL1*$^+$/*OLIG1*$^+$ program associated with m12, m1, and m4, branching into two endpoints (Figures 4F and 4G). The *ASCL1*$^+$/*OLIG1*$^+$ program was of particular interest, as it corresponded to the mGPC cluster of cells, which expressed markers associated with both astroglia (*GFAP*, *HOPX*, *EGFR*, *ASCL1*) and oligodendrocyte progenitors (*OLIG2*, *PDGFRA*), suggesting a common multipotent glial progenitor (Figures 4H and 4J). Immunohistochemistry revealed that ASCL1, OLIG2, and EGFR were often colocalized in the SVZ/IFL, oSVZ/OFL, and SP (Figures 4I, S7A, and S7B). If generated from a common glial progenitor, astrocyte and oligodendrocyte precursors might also share expression of markers associated with more differentiated states. We found that PDGFRA and OLIG2, markers associated with oligodendrocyte progenitors, and SPARCL1, which is a marker associated with mature astrocyte identity (Zhang et al., 2016), colocalized in the SVZ/IFL and oSVZ/OFL (Figures 4K and S7C–S7F). We speculate that a common multipotent glial progenitor, competent to differentiate into both astrocytes and oligodendrocytes, could explain this substantial overlap of expression programs.

Chromatin and gene expression profiles identify two astrocyte precursor populations

Human cortical astrocytes are larger, more morphologically complex (Oberheim et al., 2009; Zhang et al., 2016), and likely more diverse than those of other mammals (Vasile et al., 2017). However, the steps underlying the diversification of human astrocytes are unknown. We observed three interconnected fuzzy gene modules, largely derived from PCW24 tissue, expressing *AQP4*, *TNC*, *ALDH2*, and *APOE*, and other genes specifically expressed in astrocytes (m2, m13, m14) (Sloan et al., 2017; Wiese et al., 2012; Zhang et al., 2016) (Figures 5A, S8A, and S8B). To test whether these transcriptionally related yet distinct subpopulations associated with different regulatory factors, we computed differential motif enrichments between enhancers linked to genes in m13 versus m14. We found that the basic helix-loop-helix (bHLH) factor motifs ASCL1 and NHLH1 were enriched in module m13, while SOX21 was enriched in m14 (Figure 5B). In our glial cells, the accessibility of ASCL1 and NHLH1 motifs correlated best with the gene expression of bHLH factor *OLIG1* (Spearman rho = 0.34 and 0.36, respectively), and we have previously nominated SOX21 as a potential regulator of astrocyte maturation in cortical organoids (Trevino et al., 2020). Thus, two distinct astrocyte-like expression patterns could be distinguished by chromatin accessibility of *OLIG1* versus *SOX21* motifs.

To examine the differences between cells expressing these modules in more detail, we computed differential gene expression between the astrocytic cell clusters A1-HES and A2-OLIG, corresponding to expression of modules m2/14 and m13, respectively (Figures 5C and 5D; Table S5). Cluster A1-HES exhibited significantly higher expression of *HES4* and *CAV2*, while A2-OLIG was characterized by increased *SPARCL1*, *ID3*, and *IGFBP7* expression (Figures 5D and S8C). To determine whether these distinct astrocyte precursor subtypes were due to the sampling of different cortical areas, we used a recently published scRNA-seq dataset (Bhaduri et al., 2020) (Figures 5E and S8D). We found that gene sets attributed to our astrocytic classes were expressed in distinct cell populations in this independent dataset—an observation that could not be explained by differences in cortical area (Figure 5F). These developmental states may correspond to adult subtypes, such as protoplasmic astrocytes, found throughout the gray matter of the cortex, fibrous astrocytes found in the white matter, or primate-specific interlaminar astrocytes, which populate layer 1 (Hodge et al., 2019; Oberheim et al., 2009; Vasile et al., 2017).

Chromatin state links GPCs to lineage determination in cycling cells

We next examined how the chromatin state of progenitor cells could potentially affect the acquisition of expression programs characteristic of more differentiated cell states. We therefore focused on the heterogeneity among cells that expressed gene modules strongly associated with cell-cycle signatures (Figure 6A; Pearson r = 0.89, 0.91, respectively). To link chromatin accessibility to the glial-centric expression map, we projected pseudobulk aggregates of 13,378 glial scATAC-seq cells into our gene-module-derived manifold using accessibility-derived gene activity scores. Consistent with our CCA cluster matching analysis (Figures 2B, S3H, and S3I), pseudobulks comprised mainly of cells from ATAC cluster c15 (OPC/Oligo) projected into the oligodendrocyte endpoint of this map, cluster c10 (mGPC) data projected into the ASCL1$^+$/OLIG2$^+$ astrocyte compartment, and cluster c9 (late RG) data projected into both ependymal and HES4$^+$ astrocyte endpoints (Figure 6B). However, while we did not observe a distinct cycling cluster in our chromatin landscape, a subset of these ATAC-seq pseudobulk samples projected into the cycling, early-pseudotime compartment of the RNA-seq embedding. These samples partitioned into three distinct branches defined by their scATAC-seq cluster assignments (Figure 6C). We speculate that strong cell-cycle signatures in RNA-seq may have diminished these distinctions that are more evident in ATAC-seq data and that analyzing these

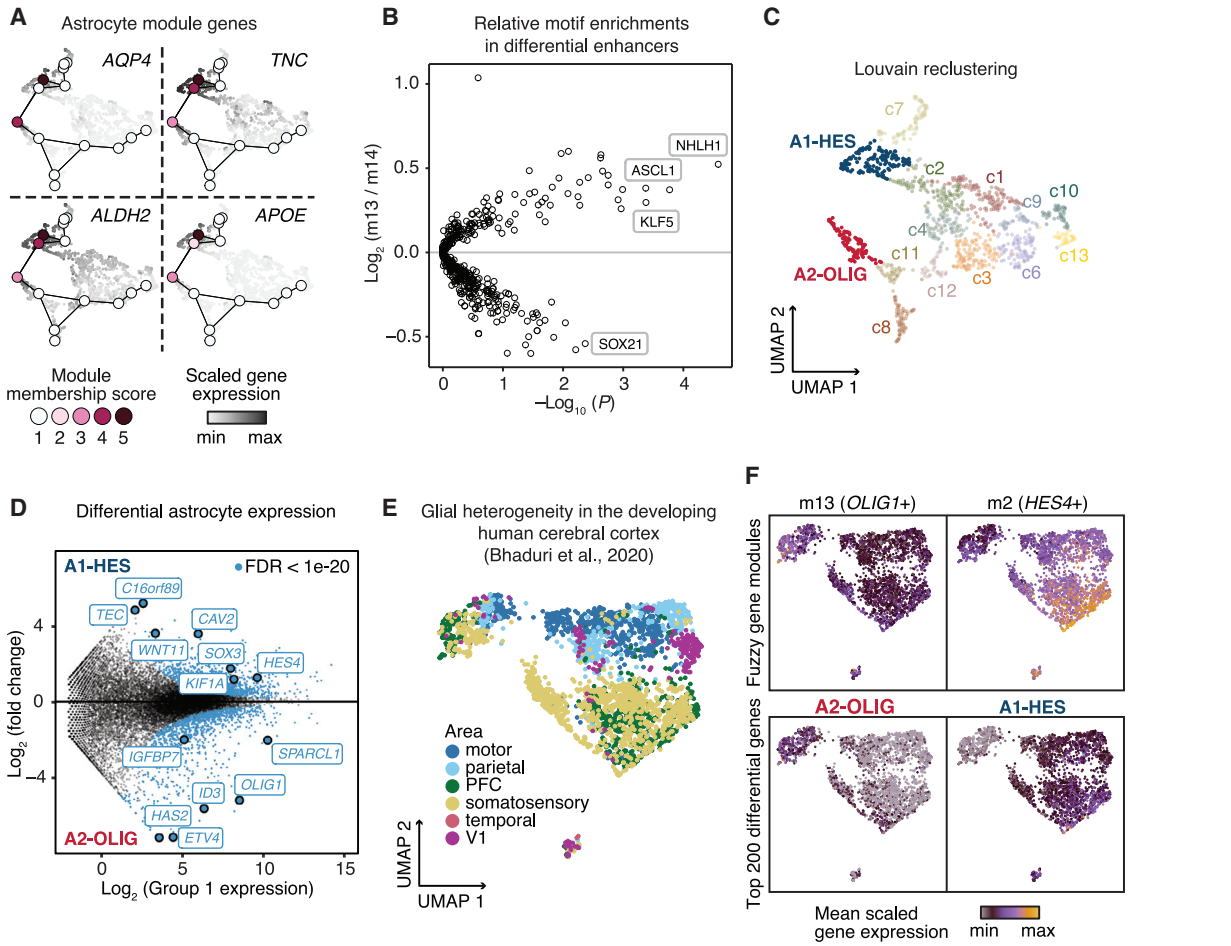


Figure 5. Astrocyte precursor heterogeneity

(A) Module membership and scaled gene expression of astrocyte-associated genes *AQP4*, *TNC*, *ALDH2*, and *APOE*

(B) Motif enrichments in GREs linked to module 13 genes relative to GREs linked to module 14.

(C) Reclustering of glial pseudobulk samples in fuzzy clustering embedding. *AQP4* positive clusters are highlighted and defined as A1-HES and A2-OLIG.

(D) Differential gene expression between A1-HES and A2-OLIG clusters, calculated using DESeq2. A threshold of Benjamini-Hochberg corrected FDR of 1e−20 was used for visualization (blue)

(E) UMAP of astrocytes from a human fetal scRNA-seq dataset (Bhaduri et al., 2020), colored by cortical area.

(F) Mean scaled expression of genes in modules m13 and m2 (top) and the top 200 differential genes from D (bottom) in Bhaduri et al. (2020).

separate branches might allow us to determine whether cycling progenitors are poised toward distinct postmitotic fates.

To explore factors that influence these fate decisions, we identified genes specific to each branch based on their gene activity scores (STAR Methods). We observed a strong overlap of these genes with the set of GPCs, including *HES1*, *RFX4*, *OLIG1*, *OLIG2*, *NEUROD6*, and *EOMES*. Overall, differential chromatin activity in all three branches of cycling cells was enriched for GPCs (Figure 6D). Each branch was enriched for at least one bHLH GPC TF in the top five most unique genes (*BHLHE40*, *OLIG1*, *OLIG2*, *NEUROD6*, *NEUROD4*) (Figure 6E). The similarity of annotated motifs for these factors is consistent with the hypothesis that they can compete for similar binding sites to drive multiple distinct cell fates (Imayoshi et al., 2013; Zhou and Anderson, 2002). Together, these results suggest that differential chromatin activity as well as gene expression of GPCs are prominent features that distinguish different types of cycling glial progenitor cells.

We next wondered whether these GPCs were both highly connected to dense collections of regulatory elements and highly enriched for lineage-defining transcription factors. To evaluate whether these links could be indicators of the eventual differentiation endpoint, and thus potentially drive differentiation, we re-projected ATAC-seq pseudobulk samples from A, B, and C cycling population branches by only using GPC-associated chromatin signals. We observed that samples moved forward in pseudotime to regions with distinct, more mature expression states (Figure 6F), whereas reprojections using random gene subsets or modules of genes moved non-specifically toward the center of the manifold (Figure S8E). This observation suggests that chromatin patterns linked to GPC genes in these cycling cells already exhibit a signature

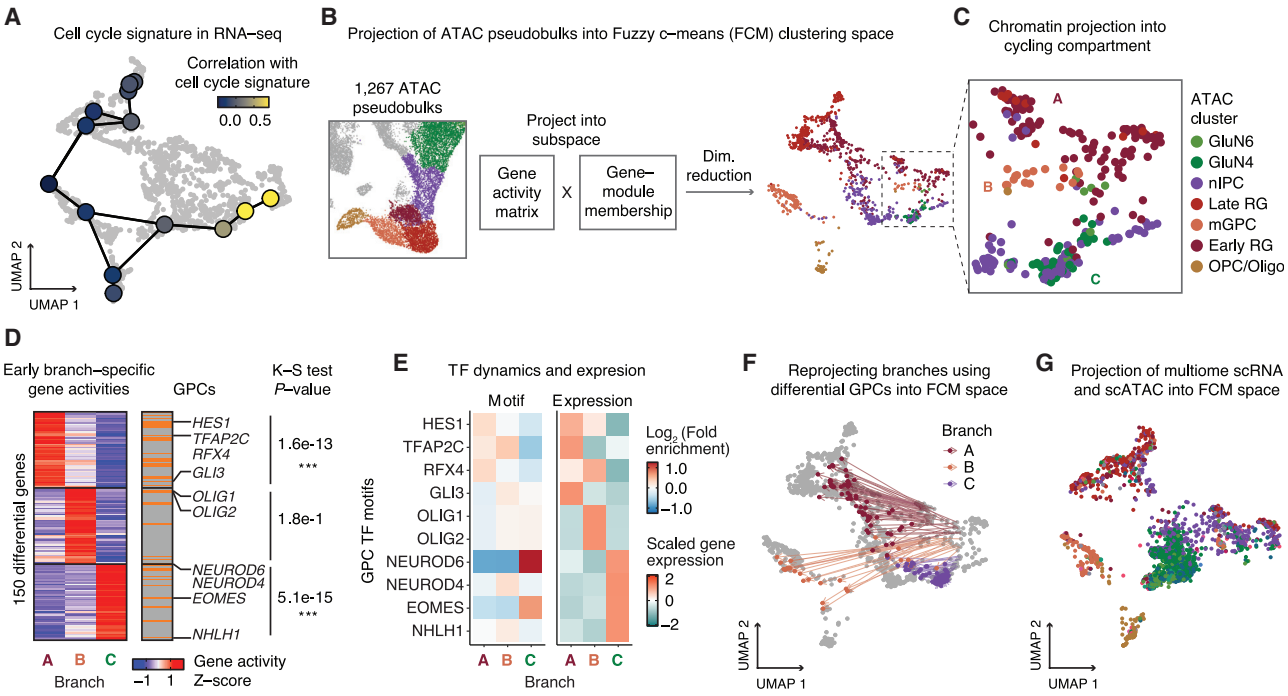


Figure 6. Chromatin state links GPCs to cell fates

(A) Pearson correlation of a cell-cycle signature (MSigDB) with module expression signature across pseudobulks.

(B) Schematic of ATAC-seq projection into fuzzy clustering embedding.

(C) Projection of ATAC-seq pseudobulks (each comprising 50 cells) into Cyc cluster and in the neighborhood of cycling-associated modules.

(D) Heatmap showing the 50 most uniquely active genes in branches A, B, and C. Gene activities are row scaled. Orange bars denote GPC labeling. The p value of a Kolmogorov-Smirnov test for enrichment of GPCs in differential, branch-specific genes are shown.

(E) Dynamics of GPC motifs and gene expression across three branches of cycling cells. Heatmaps represent enrichment of GPC TF motifs (left) and gene expression levels (right) in branch aggregates.

(F) Reprojection of branches A, B, and C using only chromatin accessibility associated with GPCs.

(G) Projection of multiome scRNA-seq data into fuzzy clustering embedding. Cells (points) are colored by their mapped scATAC-seq cluster.

of an advanced transcriptional cell state. Similarly, when we projected the scRNA-seq data from the joint multiome dataset into the module-based manifold, a fraction of cells projecting to the cycling domain exhibited distinct accessibility signatures of more differentiated cells from each branch (Figure 6G). Based on these results, we propose that during corticogenesis, progenitors entering the cell cycle may be epigenetically primed toward future cell fates and that this information is encoded specifically in GPCs, a set of genes with large numbers of linked enhancers enriched for binding of lineage-defining TFs.

Deep-learning models prioritize disruptive noncoding mutations in ASD

We next used our atlas to interpret noncoding *de novo* mutations in ASD, using the Simons Simplex Collection catalog of over 200,000 such mutations in 1,902 families (An et al., 2018) (Table S6). Naive overlap of mutations with cluster-specific scATAC-seq peaks produced no enrichment for mutations in ASD individuals relative to those in unaffected siblings (odds ratio [OR] = 1.02 for GluN6 cluster, Fisher's exact test p = 1.0; Figure S8F), indicating that peak-level annotations alone are insufficient to resolve a sparse set of causal mutations.

Deep-learning models have proven useful for prioritizing disease-relevant noncoding genetic variants based on their predicted regulatory impact (Kelley et al., 2016, 2018; Zhou and Troyanskaya, 2015). We therefore trained convolutional neural networks, based on the recent BPNet architecture, to learn models that could predict base-resolution, pseudo-bulk chromatin accessibility profiles for each of our scATAC-seq-derived cell types from genomic sequence (Figure 7A; STAR Methods) (Avsec et al., 2020), using peak regions and genomic background, matched for GC content and motif density to correct for potential sequence composition biases (Figure S8G). The models showed high and stable correlation between total predicted and observed Tn5 insertion count coverage across all peak regions in held-out chromosomes across 5-folds of cross-validated models (e.g., GluN6, mean Spearman rho = 0.58; Figure S8H; Table S6). To predict cell-context-specific effects of a candidate mutation on chromatin accessibility, we used our cluster-specific BPNet models to compute local disruption score based on the allelic fold-change in predicted counts. For each cluster, we computed the enrichment of high-effect-size mutations in cases versus controls. We observed significant enrichment of ASD-related mutations for GluN2/3/4/6/9 (>1.2-fold), which is in line with previous studies (Gandal et al.,

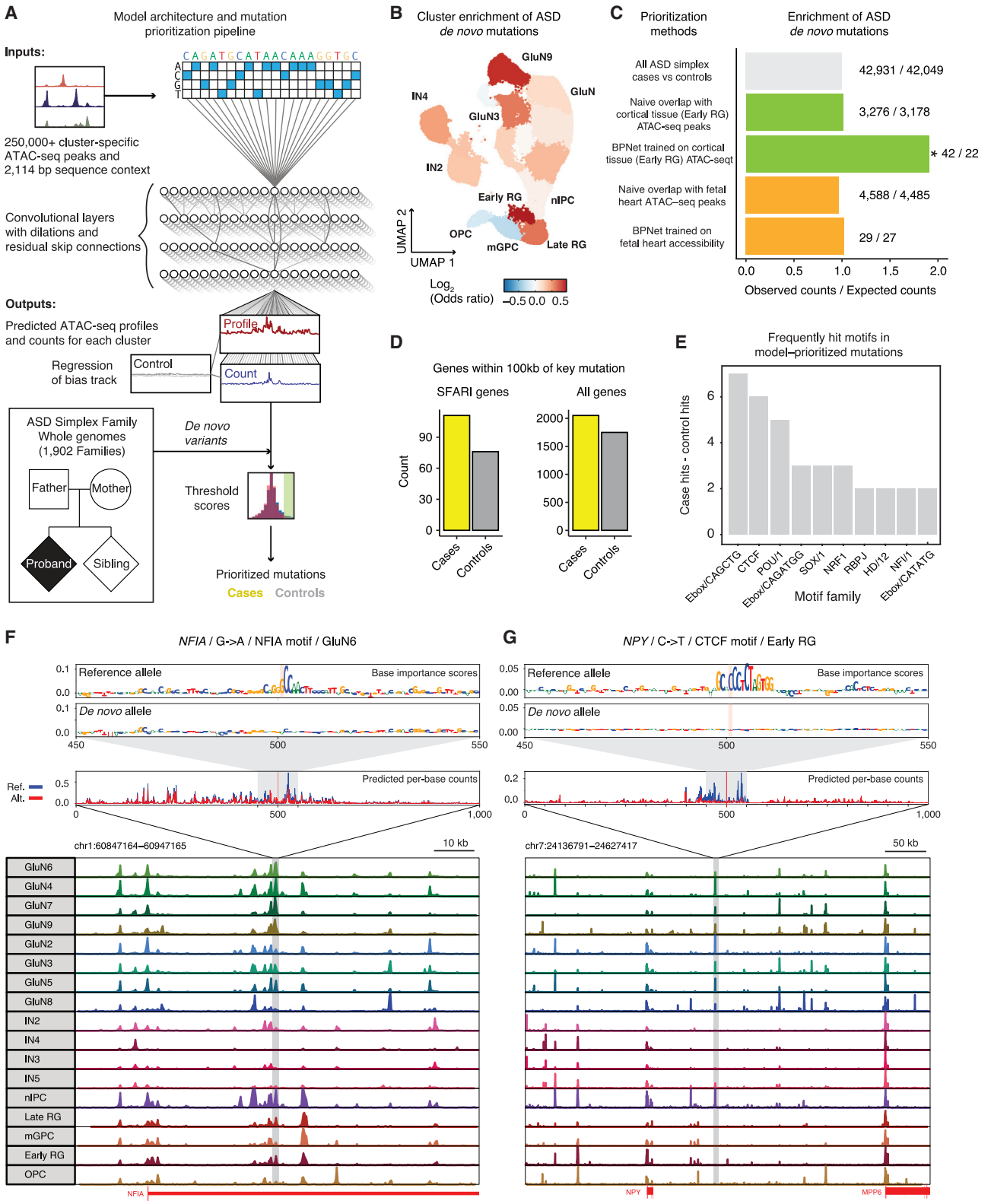


Figure 7. Disease association of gene regulatory elements

(A) Schematic of mutation prioritization pipeline.

(B) Cluster-specific BPNet enrichments visualized in scATAC UMAP.

(legend continued on next page)

2018; Li et al., 2018a; Parikshak et al., 2013; Trevino et al., 2020; Willsey et al., 2013). In addition, we found a strong association with IN2/3/4, nIPC, late RG, and early RG clusters. The early RG cluster showed the highest enrichment (OR = 1.909, excess of 20, Fisher's exact $p < 0.05$; Figure 7B; Table S6). We also observed this approach of prioritizing causal disruptive mutations was robust to threshold parameter selection (Figures S8I and S8J). In contrast, BPNet models trained on human fetal heart enhancers produced no enrichment (OR = 1.01, p = 1.0). Likewise, naive overlap enrichment with a set of fetal heart enhancers also produced no enrichment for case mutations (OR = 0.97, p = 1.0; Figure 7C). Together, these results suggest that the mutation effect scores from base-pair-resolution predictive models trained on chromatin accessibility landscapes in disease-relevant cell states are critical for prioritizing putative causal noncoding mutations.

The case and control mutations prioritized by the BPNet models had similar conservation scores and similar distances to the nearest transcription start site (TSS) (Figures S8K and S8L), highlighting the challenge of identifying these causal mutations by other means. Annotating the predicted high-effect-size mutations with their nearest genes, we observed a 1.4-fold enrichment for case mutations (n = 24) whose nearest gene was in the SFARI database compared with the control mutations (n = 17; Figure 7D). Next, we identified TF motifs that overlapped and were predicted to be disrupted by all the high-effect-size mutations from the BPNet models from all positively enriched clusters (Figure 7E, Table S6). We found that CTCF, which demarcates topological loop boundaries, was one of the most frequently disrupted motifs in cases versus controls. The NRF1 motif was another frequently disrupted motif. NRF regulates the GABA receptor subunit *GABRB1*, previously associated with disease (Li et al., 2018b). Other frequently disrupted motif families in cases relative to controls included E-box/bHLH family motifs (ASCL1, NEUROD6) and homeobox family (PAX5) motifs, with more lineage-specific effects. Homeobox proteins were also previously found to be disrupted by variants in ASD (Amiri et al., 2018; Trevino et al., 2020).

One highly disruptive mutation in our models was located in an intron of *NFIA* (Figures 7F and S8M). Loss-of-function mutations in this gene have previously been implicated in ASD (Iossifov et al., 2014). The mutation was in a linked intronic enhancer for the *NFIA* target gene. We observed that this enhancer was specifically accessible in different types of GluN clusters. The BPNet model for GluN6 predicts the mutation disrupting an NFIA motif, suggesting this mutation may dysregulate the NFIA gene expression via auto-regulatory feedback.

In the nIPC cluster, the BPNet model predicted a disruptive *de novo* mutation in an intergenic enhancer linked to the neuropeptide Y gene (*NPY*) whose TSS was 90 kb away from the mutation (Figure 7G). *NPY* is expressed in the subplate (Miller et al., 2014) and in early RG in the mid-gestation human cortex (Figure S8N), and genomic deletions of the NPY receptors have been associated with ASD (Ramanathan et al., 2004). The model further predicted this *de novo* mutation to disrupt a CTCF binding site at a chromatin loop anchor, suggesting a potential mechanistic impact on the chromatin architecture of this locus.

DISCUSSION

Here, we generate paired transcriptome and epigenome atlases of corticogenesis during a critical period of cortical development and describe how molecular interactions between DNA binding factors and *cis*-regulatory elements regulate gene expression programs. Furthermore, we describe how rare noncoding, *de novo* mutations may act to disrupt this logic.

We identified a set of genes (GPCs), enriched for lineage-defining TFs, whose local chromatin accessibility was predictive of expression levels using signals derived from single cells, possibly because of the large number of expression-linked enhancers. These linkages are evocative of other terms that have been used for similar phenomena, including "super enhancers" (Parker et al., 2013; Whyte et al., 2013) and "super-interactive promoters" (Song et al., 2020). Furthermore, chromatin accessibility of GPCs was consistent with a more differentiated cell state in some cycling progenitors. Recently, Ma et al. reported a phenomenon by which accessibility at similarly defined domains of regulatory chromatin delineate potential future cell states (Ma et al., 2020). We speculate that the coordinated effect of many enhancers on lineage-defining factors makes the expression of those factors more resistant to perturbation. Highly cooperative regulation of lineage-determining *trans*-acting factors may be a general principle of fate determination, acting as a positive feedback mechanism once a key differentiation gene has been expressed. Effectively, once activated, these enhancers might act as a ratchet, ensuring stable gene expression and preventing backtracking along a differentiation landscape when facing extrinsic or intrinsic perturbations.

Examining the trajectories of GluN migration and maturation, we found a molecular program that was consistent across 8 weeks of gestation and was defined by a sequence of motifs.

(C) Bar plot showing the enrichment of cases versus controls using different prioritization methods. Colors represent the baseline of all cleaned SSC mutations (gray), our scATAC-seq dataset (green), and a set of fetal heart enhancers (orange). * indicates a Fisher's exact test OR = 1.909, p = 0.004.
(D) Bar plot showing the number of prioritized mutations whose nearest gene is a SFARI gene. Cases (24) versus controls (17) are compared to the total number of prioritized mutations in cases (262) versus controls (232). Fisher's exact test OR = 1.24, p = 0.154.
(E) Bar plot showing the motifs that were most frequently disrupted in case mutations relative to control mutations. The y axis denotes the excess of overlaps with motifs by prioritized mutations in cases minus controls. The plot does not represent a statistical test.
(F) Example showing a disruptive case mutation in an intron of *NFIA*. The consensus logos show the importance of residues to predicted accessibility at the mutation. A 100 bp window flanking the mutation is shown. The genome tracks indicate predicted per-base counts for ref (blue) and alt (red) alleles in a 1,000 bp window flanking the mutation. The gene model around the mutation is shown along with tracks indicating the aggregate accessibility of scATAC-seq clusters at the locus.
(G) Example showing a disruptive case mutation at the *NPY* locus, as above.

Differences in neuronal regulatory activity across pseudotime were more pronounced than differences between developmental stages. We further found distinct patterns of co-accessibility and regulatory interactions between TFs early in pseudotime, whereas late TFs appeared to act more independently.

We also observed substantial sharing of TF-regulated gene expression programs among glial cells, with substantial overlap between gene modules containing canonical markers for astrocytes and oligodendrocytes. We validated the co-expression of several of these genes in human cerebral cortex. We also provided evidence for the existence of two lineages of astrocyte-like glial precursors (Vasile et al., 2017). Although glial modules were broadly interconnected, we found that the chromatin activity of GPCs in cycling cells was predictive of specific differentiated states, suggesting that progenitors entering the cell cycle are primed toward specific lineages.

Finally, our interpretable, cell-type-specific deep-learning models that link DNA sequence to chromatin accessibility can be used to assess the potential regulatory impacts of *de novo*, noncoding mutations. The modeling of the regulatory potential of individual base pairs was crucial to enable the identification of these putative causal mutations, as simple overlap with open chromatin regions did not provide the required specificity. We observed enrichments of mutations in ASD cases versus controls that approached levels observed for deleterious protein-coding mutations (An et al., 2018). We anticipate that as more large-scale ATAC-seq and RNA-seq datasets across development become available, similar approaches will allow accurate interpretation of gene-regulatory impacts of noncoding *de novo* mutations associated with other developmental disorders.

Limitations of the study

Although these data span 8 weeks of mid-gestation, an analysis at earlier and later time points would allow further study of gliogenesis and neuronal maturation and, for instance, connect astrocyte precursors to adult subtypes. Of particular interest would be to employ rapidly advancing lineage tracing methods to resolve developmental trajectories identified here. While the multiome data validate many key inferences, the use of data integration inferences to connect singleome ATAC-seq with RNA-seq and to infer lineage relationships between cells is a limitation of this study. Furthermore, our cell-specific models consider impacts of variants on peaks present only in that particular cell type. Therefore, these cell-type-specific models likely trade greater significance, afforded by scoring larger sets of overlapping mutations in pseudobulk peak calls, for a deeper understanding of the specific cell types affected by the variants. Finally, confirming the deleterious nature of noncoding *de novo* mutations prioritized in this study will require molecular validation in the cognate cell types.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human tissue and institutional approval
- METHOD DETAILS
  - Single cell dissociation
  - Single cell RNA-seq data generation
  - ATAC-seq data generation
  - Multiome data generation
  - scRNA processing
  - scATAC processing
  - Multiome data processing
  - Data analysis
  - Immunohistochemistry
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2021.07.039.

ACKNOWLEDGMENTS

We thank members of the Greenleaf, Pașca, Kundaje, and Chang labs for discussion and advice, especially X. Chen, B. Parks, F. Birey, J. Granja, and A. Banerjee. This work was supported by the Rita Allen Foundation (W.J.G.), S. Coates and the Vj Coates Foundation (S.P.P.), the Human Frontiers Science RGY006S (W.J.G), the Stanford Brain Organogenesis Program and the Big Idea Grant (S.P.P.), and the Kwan Fund (S.P.P). W.J.G. is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative. S.P.P. is a New York Stem Cell Foundation Robertson Stem Cell investigator and a Chan Zuckerberg Ben Barres investigator. H.Y.C. is an investigator of the Howard Hughes Medical Institute. Fellowship support was provided by the NSF Graduate Research Fellowship Program, the Siebel Scholars, the Enhancing Diversity in Graduate Education Program, and the Weiland Family Fellowship (A.E.T.); the Idun Berry Postdoctoral Fellowship (J.A.); the Deutsche Forschungsgemeinschaft (DFG) postdoctoral fellowship (grant MU 4303/1-1, F.M.); and the BioX Bowes Fellowship (L.S.).

AUTHOR CONTRIBUTIONS

A.E.T., F.M., J.A., S.P.P., and W.J.G. conceived the project and designed experiments. A.E.T. and F.M. performed data analysis. J.A. guided the biological interpretation of the analysis and performed validations. A.S. developed the original code (ChromBPNet) and deep-learning models of base resolution chromatin accessibility. L.S. adapted and trained the deep-learning models on the primary tissue data and performed the analysis on disease relevance with assistance from A.E.T., A.S., K.F., W.J.G., and A. Kundaje. A.M.P. and J.A. processed the samples for single-cell experiments. A.E.T. and A. Kathiria performed single-cell experiments. A.E.T., F.M., J.A., L.S., S.P.P., and W.J.G. wrote the manuscript with input from all authors. S.P.P. and W.J.G supervised the work.

DECLARATION OF INTERESTS

W.J.G. was a consultant for 10x Genomics and is named as an inventor on patents describing ATAC-seq methods. H.Y.C. is a co-founder of Accent Therapeutics and Boundless Bio and an advisor of 10x Genomics, Arsenal Biosciences, and Spring Discovery. A.S. is an employee of Insitro, Inc, and receives consulting fees from Myokardia, Inc. K.F. is an employee of Illumina, Inc.

Received: December 29, 2020
Revised: May 18, 2021
Accepted: July 28, 2021
Published: August 13, 2021

REFERENCES

Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., Menashe, I., Wadkins, T., Banerjee-Basu, S., and Packer, A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol. Autism *4*, 36.

Alexa, A., Rahnenführer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., et al. (2018). Transcriptome and epigenome landscape of human cortical development modeled in organoids. Science *362*, eaat6720.

An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2020). Deep learning at base-resolution reveals cis-regulatory motif syntax. bioRxiv. https://doi.org/10.1101/737981.

Barbarese, E., Barry, C., Chou, C.-H.J., Goldstein, D.J., Nakos, G.A., Hyde-DeRuyscher, R., Scheld, K., and Carson, J.H. (1988). Expression and localization of myelin basic protein in oligodendrocytes and transfected fibroblasts. J. Neurochem. *51*, 1737–1745.

Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., Fust, A., Preussner, J., Kuenne, C., Braun, T., et al. (2020). ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat. Commun. *11*, 4267.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414.

Bhaduri, A., Andrews, M.G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., et al. (2020). Cell stress in cortical organoids impairs molecular subtype specification. Nature *578*, 142–148.

Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. Cell *173*, 1535–1548.e16.

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, S.W., et al. (2018). The chromatin accessibility landscape of primary human cancers. Science *362*, eaav1898.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S., et al. (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309–1324.e18.

Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science *360*, eaar3131.

Fietz, S.A., Kelava, I., Vogt, J., Wilsch-Bräuninger, M., Stenzel, D., Fish, J.L., Corbeil, D., Riehn, A., Distler, W., Nitsch, R., and Huttner, W.B. (2010). OSVZ progenitors of human and ferret neocortex are epithelial-like and expand by integrin signaling. Nat. Neurosci. *13*, 690–699.

Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al.; PsychENCODE Consortium (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*, eaat8127.

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. *37*, 1458–1465.

Greig, L.C., Woodworth, M.B., Galazo, M.J., Padmanabhan, H., and Macklis, J.D. (2013). Molecular logic of neocortical projection neuron specification, development and diversity. Nat. Rev. Neurosci. *14*, 755–769.

Hansen, D.V., Lui, J.H., Parker, P.R.L., and Kriegstein, A.R. (2010). Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature *464*, 554–561.

Hodge, R.D., Bakken, T.E., Miller, J.A., Smith, K.A., Barkan, E.R., Graybuck, L.T., Close, J.L., Long, B., Johansen, N., Penn, O., et al. (2019). Conserved cell types with divergent features in human versus mouse cortex. Nature *573*, 61–68.

Imayoshi, I., Isomura, A., Harima, Y., Kawaguchi, K., Kori, H., Miyachi, H., Fujiwara, T., Ishidate, F., and Kageyama, R. (2013). Oscillatory control of factors determining multipotency and fate in mouse neural progenitors. Science *342*, 1203–1208.

Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.

Jacquet, B.V., Salinas-Mondragon, R., Liang, H., Therit, B., Buie, J.D., Dykstra, M., Campbell, K., Ostrowski, L.E., Brody, S.L., and Ghashghaei, H.T. (2009). FoxJ1-dependent gene expression is required for differentiation of radial glia into ependymal cells and a subset of astrocytes in the postnatal brain. Development *136*, 4021–4031.

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–489.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

Kelley, D.R., Snoek, J., and Rinn, J.L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. *26*, 990–999.

Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. *28*, 739–750.

Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. Science *358*, 69–75.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G., et al. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. *46* (D1), D260–D266.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. (2019). Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. *20*, 207–220.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Li, M., Santpere, G., Kawasawa, Y.I., Evgrafov, O.V., Gulden, F.O., Pochareddy, S., Sunkin, S.M., Li, Z., Shin, Y., Zhu, Y., et al. (2018a). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science *362*, eaat7615.

Li, Z., Cogswell, M., Hixson, K., Brooks-Kayal, A.R., and Russek, S.J. (2018b). Nuclear Respiratory Factor 1 (NRF-1) Controls the Activity Dependent Transcription of the GABA-A Receptor Beta 1 Subunit Gene in Neurons. Front. Mol. Neurosci. *11*, 285.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and evolution of the human neocortex. Cell *146*, 18–36.

Lundberg, S., and Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. arXiv, 1705.07874. https://arxiv.org/abs/1705.07874.

Ma, S., Zhang, B., LaFave, L., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103–1116.e20.

McConnell, S.K. (1995). Constructing the cerebral cortex: neurogenesis and fate determination. Neuron *15*, 761–768.

McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329–337.e4.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv, 1802.03426. https://arxiv.org/abs/1802.03426.

Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206.

Molnár, Z., Clowry, G.J., Šestan, N., Alzu'bi, A., Bakken, T., Hevner, R.F., Hüppi, P.S., Kostović, I., Rakic, P., Anton, E.S., et al. (2019). New insights into the development of the human cerebral cortex. J. Anat. *235*, 432–451.

Nowakowski, T.J., Pollen, A.A., Sandoval-Espinosa, C., and Kriegstein, A.R. (2016). Transformation of the Radial Glia Scaffold Demarcates Two Stages of Human Cerebral Cortex Development. Neuron *91*, 1219–1227.

Nowakowski, T.J., Bhaduri, A., Pollen, A.A., Alvarado, B., Mostajo-Radji, M.A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S.J., Velmeshev, D., et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science *358*, 1318–1323.

Oberheim, N.A., Takano, T., Han, X., He, W., Lin, J.H.C., Wang, F., Xu, Q., Wyatt, J.D., Pilcher, W., Ojemann, J.G., et al. (2009). Uniquely hominid features of adult human astrocytes. J. Neurosci. *29*, 3276–3287.

Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008–1021.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J.A., van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al.; NISC Comparative Sequencing Program; National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. USA *110*, 17921–17926.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., et al. (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell *71*, 858–871.e8.

Polioudakis, D., de la Torre-Ubieta, L., Langerman, J., Elkins, A.G., Shi, X., Stein, J.L., Vuong, C.K., Nichterwitz, S., Gevorgian, M., Opland, C.K., et al. (2019). A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron *103*, 785–801.e8.

Pollen, A.A., Nowakowski, T.J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C.R., Shuga, J., Liu, S.J., Oldham, M.C., Diaz, A., et al. (2015). Molecular identity of human outer radial glia during cortical development. Cell *163*, 55–67.

Ramanathan, S., Woodroffe, A., Flodman, P.L., Mays, L.Z., Hanouni, M., Modahl, C.B., Steinberg-Epstein, R., Bocian, M.E., Spence, M.A., and Smith, M. (2004). A case of autism with an interstitial deletion on 4q leading to hemizygosity for genes encoding for glutamine and glycine neurotransmitter receptor sub-units (AMPA 2, GLRA3, GLRB) and neuropeptide receptors NPY1R, NPY5R. BMC Med. Genet. *5*, 10.

Ransom, B.R. (2012). Neuroglia (Oxford University Press).

Rubenstein, J.L.R. (2011). Annual Research Review: Development of the cerebral cortex: implications for neurodevelopmental disorders. J. Child Psychol. Psychiatry *52*, 339–355.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978.

Sheffield, N.C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics *32*, 587–589.

Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning Important Features Through Propagating Activation Differences. arXiv, 1704.02685. https://arxiv.org/abs/1704.02685.

Silbereis, J.C., Pochareddy, S., Zhu, Y., Li, M., and Sestan, N. (2016). The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. Neuron *89*, 248–268.

Sloan, S.A., Darmanis, S., Huber, N., Khan, T.A., Birey, F., Caneda, C., Reimer, R., Quake, S.R., Barres, B.A., and Paşca, S.P. (2017). Human Astrocyte Maturation Captured in 3D Cerebral Cortical Spheroids Derived from Pluripotent Stem Cells. Neuron *95*, 779–790.e6.

Smit, A.F.A., Hubley, R., and Green, P. (2010). RepeatMasker Open-3.0. http://www.repeatmasker.org.

Song, M., Pebworth, M.-P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J.D., Choudhary, M.N.K., Cui, X., Jones, I.R., et al. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. Nature *587*, 644–649.

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S.L., Vernot, B., Cheng, J.B., Thurman, R.E., Sandstrom, R., et al. (2013). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell *154*, 888–903.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Thomsen, E.R., Mich, J.K., Yao, Z., Hodge, R.D., Doyle, A.M., Jang, S., Shehata, S.I., Nelson, A.M., Shapovalova, N.V., Levi, B.P., and Ramanathan, S. (2016). Fixed single-cell transcriptomic characterization of human radial glial diversity. Nat. Methods *13*, 87–93.

Tjärnberg, A., Mahmood, O., Jackson, C.A., Saldi, G.-A., Cho, K., Christiaen, L.A., and Bonneau, R.A. (2021). Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data. PLoS Comput. Biol. *17*, e1008569.

Trevino, A.E., Sinnott-Armstrong, N., Andersen, J., Yoon, S.-J., Huber, N., Pritchard, J.K., Chang, H.Y., Greenleaf, W.J., and Paşca, S.P. (2020). Chromatin accessibility dynamics in a model of human forebrain development. Science *367*, eaay1645.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716–729.e27.

Vasile, F., Dossi, E., and Rouach, N. (2017). Human astrocytes: structure and functions in the healthy brain. Brain Struct. Funct. *222*, 2017–2029.

Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wiese, S., Karus, M., and Faissner, A. (2012). Astrocytes as a source for extracellular matrix molecules and cytokines. Front. Pharmacol. *3*, 120.

Willsey, A.J., Sanders, S.J., Li, M., Dong, S., Tebbenkamp, A.T., Muhle, R.A., Reilly, S.K., Lin, L., Fertuzinhos, S., Miller, J.A., et al. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell *155*, 997–1007.

Wonders, C.P., and Anderson, S.A. (2006). The origin and specification of cortical interneurons. Nat. Rev. Neurosci. *7*, 687–696.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., et al. (2016). Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. Neuron *89*, 37–53.

Zhou, Q., and Anderson, D.J. (2002). The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. Cell *109*, 61–73.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat. Genet. *51*, 973–980.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse anti-ASCL1	BD Biosciences	556604
Rat anti-CTIP2	Abcam	ab18465
Rat anti-EGFR	Abcam	ab231
Rabbit anti-GFAP	Dako	Z0334
Rat anti-GFAP	Thermo Fisher Scientific	13-0300
Mouse anti-HOPX	Santa Cruz	sc-398703
Mouse anti-KI67	BD Biosciences	550609
Rabbit anti-OLIG2	Millipore	AB9610
Rabbit anti-PBXIP1	Abcam	ab84752
Rabbit anti-PDGFRA	Santa Cruz	sc-338
Rabbit anti-PPP1R17	Atlas Antibodies	HPA047819
Goat anti-SOX9	R&D Systems	AF3075
Goat anti-SPARCL1	Novus Biologicals	AF2728
Rabbit anti-TFAP2C	Thermo Fisher Scientific	14572-1
Critical commercial assays		
Chromium Single cell 3′ GEM, Library & Gel Bead Kit v3	10x Genomics	PN: 1000075
ATAC-seq NextGEM kit	10x Genomics	PN: 1000175
Single Cell Multiome ATAC + Gene Expression kit	10x Genomics	PN: 1000285
Deposited data		
scRNaseq, scATACseq and multiome data	This study	GEO: GSE162170
Software and algorithms		
ImageJ (Fiji)		https://imagej.net/software/fiji
Genome assembly	https://www.ncbi.nlm.nih.gov/grc/human	hg38 / GRCh38
Gene annotation	https://www.gencodegenes.org/human/release_27.html	Gencode v.27
Cell Ranger	10x Genomics	CellRanger v3.1.0
Cell Ranger-ATAC	10x Genomics	Cell Ranger-ATAC v1.2.0
Cell Ranger-ARC	10x Genomics	Cell Ranger-ARC v1.0.0
Seurat	Stuart et al., 2019; Bioconductor	Seurat v.3.1.4
Velocyto	La Manno et al., 2018	Velocyto v.0.17.17
scVelo	Bergen et al., 2020	scVelo v.0.1.2
DoubletFinder	McGinnis et al., 2019	DoubletFinder v.2.0.2
MACS2	Zhang et al., 2008	MACS2 v2.1.1
GenomicRanges	Bioconductor	GenomicRanges v.1.36
ChrAccR	https://greenleaflab.github.io/ChrAccR/	ChrAccR v.dev.0.9.11+
ChromVAR	Schep et al., 2017	ChromVAR v.1.6
Motifmatchr	Bioconductor	motifmatchr_1.6.0
JASPAR TF motifs	Khan et al., 2018	JASPAR 2018
TF motif clusters	Vierstra et al., 2020; https://www.vierstra.org/resources/motif_clustering	Motif-clustering v1.0

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MAGIC	van Dijk et al., 2018	N/A
Uwot	McInnes et al., 2020	uwot v.0.1.8
DESeq2	Love et al., 2014; Bioconductor	DESeq2_1.24.0
scRNaseq data from developing human cerebral cortex	Bhaduri et al., 2020; https://cells.ucsc.edu	Dataset ID: 'organoidreportcard/primary10X'
scRNaseq data from the Allen Brain Atlas	https://portal.brain-map.org/atlases-and-data/rnaseq	MTG - SMART-SEQ Hodge et al., 2019
scRNA-seq data	Polioudakis et al., 2019	https://geschwindlab.dgsom.ucla.edu/pages/codexviewer
URD	Farrell et al., 2018	URD v.1.1.1
GenomicInteractions	Bioconductor	GenomicInteractions_1.20.3
LOLA	Sheffield and Bock, 2016; Bioconductor	LOLA v.1.14.0
topGO	Alexa et al., 2006	topGO v.2.36.0
GenomicScores	Bioconductor	GenomicScores_1.10.0
Cmeans	CRAN	e1071_1.7-3
clusterProfiler	Yu et al., 2012; Bioconductor	clusterProfiler v.4.0.2
Astrocyte genes	Zhang et al., 2016	supplemental tables from Zhang et al.
De novo mutations from the Simons Simplex Collection	An et al., 2018	supplemental materials Table S2 from An et al.
De novo mutation calls in gnomAD	Karczewski et al., 2020	supplemental information: variant annotation from Karczewski et al.
RepeatMasker	Smit et al., 2010	http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/rmsk.txt.gz
Shap.deep_explainer	Lundberg and Lee, 2017	https://shap-lrjball.readthedocs.io/en/latest/generated/shap.DeepExplainer.html
DeepLift	Shrikumar et al., 2019	N/A
Other		
Papain enzyme solution	Worthington	LS03126
DNase	Worthington	LS002007
Digitonin	Promega	G9441
RNase inhibitor	Roche	3335402001

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact Sergiu P. Pașca (spasca@stanford.edu)

Materials availability

This study did not generate unique reagents.

Data and code availability

- Data used for the analyses presented in this work are available under GEO: GSE162170 and on a supplementary website (see below).
- BPNet code can be found at: https://github.com/GreenleafLab/Brain_ASD. A supplementary website with references to the data, code repositories and tools for interactive data exploration (cell browser and genome browser tracks) can be found at https://scbrainregulation.su.domains/.
- Please see the supplementary materials for detailed availability of the provided tables, and data.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human tissue and institutional approval

De-identified tissue samples were obtained at Stanford University School of Medicine from elective pregnancy terminations under a protocol approved by the Research Compliance Office at Stanford University. Brain tissue samples at PCW16 to PCW24 were delivered on ice and processed for single cell analyses or immunocytochemistry within 3 h of the procedure.

METHOD DETAILS

Single cell dissociation

Dissociation of human tissue into single cells was performed as described with some modifications (Sloan et al., 2017; Trevino et al., 2020). Briefly, tissue was chopped and incubated in 30 U/mL papain enzyme solution (Worthington, LS03126) and 0.4% DNase (12,500 units/mL; Worthington, LS002007) at 37°C for 45 min. After digestion, samples were washed with a protease inhibitor solution and gently triturated to achieve a single cell suspension.

Single cell RNA-seq data generation

Cells were resuspended in 0.02% BSA/PBS and passed through a 70 μm filter before proceeding to single-cell sample preparation. Single-cell libraries were prepared using the RNA 3′ v3 protocol (10x Genomics), loading 7,000 cells per lane.

ATAC-seq data generation

For ATAC-seq, nuclei were prepared on ice or in a centrifuge at 4°C. All centrifugation steps were run for 5 min at 500 x g. 100,000 dissociated cells were washed in ice-cold ATAC-seq resuspension buffer (RSB, 10 mM Tris pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$), spun down, and resuspended in 100 μL ATAC-seq lysis buffer (RSB plus 0.1% NP-40 and 0.1% Tween-20 (Thermo Fisher). Lysis was allowed to proceed on ice for 5 min, then 900 μL RSB was added before spinning down again and resuspending in 50 μL 1X Nuclei Resuspension Buffer (10x Genomics). A sample of the nuclei was stained with Trypan Blue and inspected to confirm complete lysis. If necessary, cell concentrations were adjusted prior to starting single-cell droplet generation with the ATAC-seq NextGEM kit (10x Genomics). 4,000 nuclei were loaded per lane.

Multiome data generation

For multiome single cell data, nuclei were prepared as above for ATAC-seq with minor changes. Specifically, 0.01% digitonin was added to the lysis buffer, and 2 U/μL RNase inhibitor (Roche) was added to all nuclei preparation buffers. After nuclei preparation, droplets and single cell libraries were prepared using the Single Cell Multiome ATAC + Gene Expression kit (10x Genomics) and 4,000 nuclei were loaded per lane.

scRNA processing

Raw sequencing data were converted to fastq format using the command 'cellranger mkfastq' (10x Genomics, v.3.1.0). scRNA-seq reads were aligned to the GRCh38 (hg38) reference genome and quantified using 'cellranger count' (10x Genomics, v.3.1.0). 'Velocyto' (v.0.17.17) (La Manno et al., 2018) was used to obtain splicing-specific count data for downstream RNA velocity analysis.

Count data was further processed using the 'Seurat' R package (v.3.1.4) (Stuart et al., 2019), using Gencode v.27 for gene identification. We removed cells with less than 500 informative genes expressed, cells with less than 500 sequenced fragments and cells with more than 40% of counts corresponding to mitochondrial genes. Genes not contained in the Gencode annotation were excluded from further analysis. We performed doublet analysis using the 'DoubletFinder' R package (v.2.0.2) (McGinnis et al., 2019), but did not find clear evidence of cell doublets biasing our unsupervised analysis and therefore did not apply doublet filtering. Count data was log-normalized and scaled to 10,000. PCA analysis was based on the 2,000 most variable genes. The top 50 principal components (PCs) were retained for further analysis, excluding one component because it was strongly associated with the expression of more than 5 genes related to cell stress (*HSPA*, *JUN*, *FOS*, *DUSP* gene families). Nearest neighbors were computed based on the PC representation, and 23 clusters were identified using Louvain clustering implemented in Seurat's 'FindClusters' function ('resolution = 0.5'). 2-dimensional representations were generated using uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) as implemented in Seurat and the 'uwot' R packages (v.0.1.8; parameter settings: 'min.dist = 0.8', 'n.neighbors = 50', 'cosine' distance metric).

scATAC processing

Raw sequencing data were converted to fastq format using 'cellranger-atac mkfastq' (10x Genomics, v.1.2.0). scATAC-seq reads were aligned to the GRCh38 (hg38) reference genome and quantified using 'cellranger-atac count' (10x Genomics, v.1.2.0).

Fragment data was further processed using the 'ChrAccR' R package (v.dev.0.9.11+). We filtered out cells with less than 1,000 or more than 50,000 sequencing fragments. TSS enrichment was computed as a metric of signal-to-noise ratio using methods described in (Granja et al., 2019) and we discarded cells with a TSS enrichment less than 4. Fragments on sex chromosomes and mitochondrial DNA were excluded from downstream analysis.

In order to obtain a low dimensional representation of single-cell ATAC datasets in terms of principal components and UMAP coordinates, we applied an iterative latent semantic indexing approach (Granja et al., 2019). This approach also identified 22 cell clusters and a consensus set of 657,930 cluster peaks. In brief, in an initial iteration clusters were identified based on the 20,000 most accessible 5kb-tiling regions. Here, the counts were first normalized using the term frequency - inverse document frequency (TF-IDF) transformation (Cusanovich et al., 2018), and singular values were computed based on these normalized counts. Initial clusters were identified based on the top 25 singular values using Louvain clustering (as implemented in the Seurat package, resolution parameter = 0.6), excluding the first singular value as it exceeded a correlation coefficient of 0.5 with read depth. Peak calling was then performed on the aggregated insertion sites from all cells of each cluster using MACS2 (v2.1.1). A consensus set of peaks uniform-length non-overlapping peaks was obtained by selecting the peak with highest score from each set of overlapping peaks. In a second iteration, the 50,000 peaks whose TF-IDF-normalized counts exhibited the highest variability across the initial clusters provide the basis for a refined clustering using the top 50 derived singular values. In the final iteration, the 50,000 most variable peaks across the refined clusters were identified as the final peak set and singular values were computed again. UMAP coordinates and ATAC clusters were determined based on the top 10 of these final singular values. 2-dimensional representations were generated using UMAP as implemented in the 'uwot' R package (v.0.1.8; parameter settings: 'min.dist = 0.6', 'n.neighbors = 50', 'cosine' distance metric).

ChromVAR (Schep et al., 2017) (v.1.6) was used to obtain TF accessibility profiles using position weight matrices from the JASPAR 2018 database (Khan et al., 2018). Gene activity scores were computed as the aggregated accessibility of TSS-associated peaks using 'ChrAccR'. For this, counts in peaks within 100,000 bp of a TSS have been summed up using weights assigned by a radial basis function (RBF) with a width parameter sigma = 10,000 bp, setting a minimum asymptotic weight of 0.25. For each gene, the resulting scores were normalized by the sum of the weights. For visualization and downstream analysis counts from single-cells have been rescaled to 10,000 counts and have been $\log_2$-normalized. For enhanced visualization in 2-dimensional UMAP space, gene activity scores have been smoothed using the MAGIC diffusion algorithm (van Dijk et al., 2018) with cell neighborhoods determined in singular value space. Since such imputation methods have been associated with the risk of oversmoothing (Tjärnberg et al., 2021), we restricted the application of MAGIC to data visualization.

We created ATAC signal tracks by summing insertion counts in cluster pseudobulk samples in 200bp genomic tiling windows and provide trackhub compatible with the WashU Epigenome Browser (http://epigenomegateway.wustl.edu) containing these profiles in addition to inferred CRE-gene links.

Multiome data processing

Raw sequencing data were converted to fastq format using 'cellranger-arc mkfastq' (10x Genomics, v.1.0.0). scATAC-seq reads were aligned to the GRCh38 (hg38) reference genome and quantified using 'cellranger -arc count' (10x Genomics, v.1.0.0).

RNA count data was further processed using 'Seurat' as described above, with the exception that all 50 principal components were retained. This resulted in 9,818 cells after filtering, which were assigned to 14 clusters in the unsupervised analysis. ATAC fragment data was further processed using 'ChrAccR' as described above, resulting in 9,091 cells post-filtering, assigned to 16 clusters and a consensus peak set of 467,315 elements. Jointly applying ATAC and RNA filters resulted in 8,981 cells with high-quality measurements across both modalities.

Data analysis

Matching of single-cell transcriptomes and epigenomes

Canonical correlation analysis (CCA) as implemented in Seurat has been applied to matched single-cell RNA and ATAC data from each gestational time point individually. For this purpose, we computed log-normalized and scaled gene activity scores as surrogates for gene expression in the cells profiled by scATAC-seq. As integration features, we used the union of the 2,000 most variable genes in each modality as input to Seurat's 'FindTransferAnchors' function with reduction method 'cca' and parameter 'k.anchor = 10'. For each cell profiled by scRNA-seq and each cell profiled by scATAC-seq we identified the nearest neighbor cell in the respective other modality by applying nearest-neighbor search in the joint CCA L2 space. Nearest neighbors were determined using the 'FNN' R package employing the 'kd_tree' algorithm with Euclidean distance. These nearest-neighbor-based cell matches from all gestational time points were concatenated to obtain dataset-wide cell matches across both modalities.

Linking gene regulatory elements and gene expression across all cell types

We identified peak-to-gene links using a correlation-based approach (Corces et al., 2018) applied to pseudobulk samples aggregating scATAC and scRNA counts. These pseudobulk samples were defined by randomly sampling 200 cells from the entire scATAC-seq dataset. These 200 seed cells were combined with their respective 99 nearest neighbor cells in ATAC-PC space, such that each pseudobulk sample comprised 100 cells in total. Pseudobulk ATAC insertion counts for peaks were obtained by summing peak insertion counts across the respective single-cell members. Matching RNA cells were obtained by selecting the 100 scRNA cells that resembled nearest neighbors to the 100 ATAC cells in CCA space. Pseudobulk RNA gene counts were obtained by summing gene counts across the respective single-cell members. Similarly, in the multiome dataset, 200 pseudobulk samples of 100 cells each were sampled from the ATAC modalitity, and the same cells were aggregated in RNA space. Each matched pseudobulk sample was annotated with the majority cluster and age assignments of its contingent RNA and ATAC cells respectively.

We then obtained candidate peak-gene pairs by associating peaks with a genomic distance between 1 and 250 kb to the TSS of protein coding and lincRNA genes to the respective genes. For each candidate peak-gene pair we computed the Pearson correlation coefficient of CPM-normalized counts of accessibility and gene expression data and computed FDR-adjusted *P*-values for these coefficients based on their *t*-statistic. We defined a set of 64,878 high-confidence peak-to-gene links by only retaining pairs with |PCC| >0.4 and FDR-adjusted *P*-value < 0.05. Using the same method, a corresponding set of 76,374 links was obtained for the multiome data. Overlap between inferred and multiome peak-gene links was computed by creating ''GenomicInteraction'' objects for each, with the peak as the first anchor and the gene promoter as the second, then applying the function 'findOverlaps' with parameter ''use.region = 'both'.''

Validation of inferred peak-gene links using conservation and chromosome conformation capture data

To validate the linkages identified above using orthogonal analyses, two approaches were taken. First, for multiome and singleome linkages, phastCons 100-way vertebrate conservation scores were computed for linked and unlinked peaks using the ''gscores'' function from the ''GenomicScores'' package. Scores for linked and unliked peaks were compared using a Wilcoxon rank-sum test.

Second, we performed an analysis of 3D contacts using a recently published proximity ligation assisted ChIP sequencing (PLAC-seq) dataset targeting H3K4me3 sites in 4 sorted developmental human cell types from cerebral cortex (Song et al., 2020). The dataset was comprised of promoter capture 3D contact libraries from FACS sorted interneurons, excitatory neurons, radial glia, and intermediate progenitor cells from dissociated human cerebral cortical tissue. Thus, this orthogonal 3D contact dataset provided two axes of validation: first, the enhancer-promoter linkages, and second, the cell-type specificity of these linkages.

Interaction calls from PLAC-seq data were imported as ''GenomicInteractions'' objects and overlapped with our linkages (''findOverlaps''). For validation, both anchors of of both interactions were required to overlap. We performed this analysis for all possible peak-gene links, significant inferred peak-gene links, and, since these orthogonal data types do not correspond 1:1, for an independent test we also pre-filtered our significant links for any 1-dimensional interaction with PLAC-seq regions prior to the overlap analysis.

To account for the skewed length distribution of significant links, we also generated 1000 length-matched permutations from the space of all possible links (10,000 links each). First, for significant peak-gene links, the peak-promoter distance was computed. Distances were binned into 25 equal bins from 0-250 kb, and the proportion of peak-gene links in each bin was calculated. Next, we assigned all possible peak-gene links a bin and the corresponding proportion from the true distribution. The proportions were used as sampling probabilities for drawing the permutations. Then, PLAC-seq overlaps were computed against this length-matched null model (second bar in Figure S3K).

Finally, we reasoned that, if the inferred linkages were valid, validated genes would also exhibit cell-type restricted expression patterns, in line with the sorted cell types of the 3D contact data. To determine this, we computed the expression of linked genes in the major cell types in our RNA-seq data. Then we partitioned those expression values by the cell type origin of the PLAC-seq interaction. Similarly, for linked ATAC-seq peaks, we computed the mean accessibility in scATAC-seq data over the same margins.

Projection of external datasets into the scRNA landscape

We retrieved scRNA data from the developing human cerebral cortex (Bhaduri et al., 2020). We downloaded the normalized data from the UCSC Cell Browser (https://cells.ucsc.edu; dataset ID: 'organoidreportcard/primary10X') and the data was read into a Seurat object using custom R scripts. We then projected the data into our scRNA UMAP space using the 'uwot' model stored in our dataset, i.e., we used an identical principal component gene loadings and 'uwot' model parametrization. This UMAP space representation allowed us to assign a nearest neighbor from our scRNA cells to each cell in the external dataset. Cell annotation (pseudotime, cell cluster, etc.) were transferred from these nearest neighbors. Jaccard indices were computed between the transferred annotation and the downloaded external metadata.

Similarly, we downloaded 10x Genomics scRNA data from the Allen Brain Map (https://portal.brain-map.org/atlases-and-data/rnaseq). The downloaded raw count data was read into a Seurat object and processed using the same steps and parameters used for processing our scRNA data. Projection and annotation transfer were done in the same way as for the external developing brain dataset. For Figures S4E–S4G, we restricted the projection to cells labeled as excitatory neurons ('Exc') in the external cell metadata.

Third, we obtained scRNA-seq data from the developing human neocortex at mid-gestation (PCW17-18) (Polioudakis et al., 2019). We downloaded count matrix and cell annotation data from https://geschwindlab.dgsom.ucla.edu/pages/codexviewer. Projection, annotation transfer and computation of Jaccard correspondences were done as for the other external datasets.

Projection of multiome data into the scRNA and scATAC landscapes

Based on the RNA-based gene counts, we projected the multiome data into our scRNA UMAP space using the 'uwot' model stored in our scRNA dataset, i.e., we used an identical principal component gene loadings and 'uwot' model parametrization. Similarly, multiome cells were projected into scATAC UMAP space based on the 'uwot' model derived from the scATAC dataset using the same peak loadings. We used these projections to assign a nearest neighbor from our scRNA cells or scATAC cells to each cell in the multiome dataset. Cell annotation (pseudotime, cell cluster, etc.) were transferred from these nearest neighbors.

Identification of genes with predictive chromatin (GPCs)

The definition of GPCs is primarily based on high gene activity-expression correlations across single cells. To make this analysis more robust to technical variation, we restricted our analysis to the most variable genes across dorsal forebrain cells (1999 genes). Specifically, we used the ''findVariableGenes'' function from the URD package with parameters ''diffCV.cutoff = 0.15, mean.min = 0.004''

(Farrell et al., 2018). For each variable gene, we computed Spearman's correlation coefficients between the vector of gene activity scores for ATAC cells and the vector of expression scores in the corresponding nearest neighbor cells in RNA data. We also compared these correlations to the number of linked enhancers per gene (see above). From this subset, we defined GPCs as genes in the top 10% of gene activity-expression correlations that were linked to a minimum of 10 CREs.

Definition of RNA velocity and pseudotime in excitatory neuron trajectories

Excitatory neuron trajectories were defined based on RNA cells in selected clusters (*cf.* Table S1). We computed RNA velocity using custom R scripts interfacing with the 'scVelo' toolkit (v.0.1.25) (Bergen et al., 2020) via the 'reticulate' R-Python interface. For this, we exported the Velocyto-derived spliced and unspliced counts along with Seurat-derived PC and UMAP representations of single cells as 'AnnData' objects. We filtered the dataset using the scVelo function 'pp.filter_and_normalize' (parameters: min_shared_counts = 10, n_top_genes = 2,000) and computed moments using 'pp.moments' (n_pcs = 30, n_neighbors = 30). We then used 'tl.velocity' with mode = 'stochastic' to compute cell velocities and 'tt.velocity_graph' to compute a velocity graph. Potential root and end point cells for the trajectory were computed using 'tt.terminal_states'. To compute cell pseudotime scores, we employed a modified version of the scVelo function 'tt.velocity_pseudotime'. In contrast to the original version of the function which combines diffusion estimates from a forward pass starting in the root cells and a backward pass starting in the end point cells, this modified version only applies the forward pass starting in the root cells. This was necessary because scVelo-identified end point cells that were inconsistent with our notion of trajectory. We re-imported the scVelo-derived cell annotations (velocity vectors, pseudotime, root and end point probabilities) into the metadata of the R-based Seurat objects. Finally, cell pseudotime scores were rescaled to their quantiles using the R function 'ecdf'.

Additionally, in order to quantify when in pseudotime a gene is expressed we computed a weighted average pseudotime value. We define this 'gene pseudotime' for each gene j as

$$\tau_j = \sum_{i=1}^{N} t_i \frac{c_{ij}}{\sum_{k=1}^{N} c_{kj}},$$

where N = *363* is the number pseudobulk samples used for linking regulatory elements to genes (see below), t_i is the mean pseudotime across all cells in pseudobulk sample i, and c_{ij} is the aggregate RNA count for pseudobulk sample i in gene j.

Pseudotime of cells profiled using scATAC-seq were defined as the pseudotime of their nearest RNA-cell neighbor in CCA space.

Linking gene regulatory elements and gene expression in the excitatory neuron trajectory

To facilitate aggregate analysis along pseudotime, we obtained pseudobulk samples by sorting cells based on their pseudotime scores and merging bins of 100 cells. The same correlation-based approach as used on all cell types (see above) was applied to these pseudobulk samples, linking peaks to cluster-specific genes. These cluster-specific genes were identified from the scRNA data of cells included in the excitatory neuron trajectory employing a Wilcoxon test as implemented in Seurat's 'FindAllMarkers' function and applying thresholds of 0.01 and 0 for test-derived *P*-values and log(fold-changes) respectively. We retained links with accessibility-gene expression correlation coefficients with PCC > 0.4 and FDR-adjusted *P*-value < 0.05, which resulted in 13,989 high-confidence positively correlated peak-to-gene links with specificity to the excitatory neuron trajectory. These links were clusters using k-means (k = 5) clustering based on the z-score-scaled expression levels of the associated genes. Enrichment analysis for these clusters were performed using the 'topGO' (v.2.36.0) R/Bioconductor package (Gene Ontology enrichment), Fisher's exact tests on manually curated gene sets and Fisher's exact tests as implemented in the R/Bioconductor package 'LOLA' (Sheffield and Bock, 2016) (v.1.14.0) for peak TF motif occurrences (based on genome-wide scans of JASPAR 2018 PWMs using the 'motifmatchr' R package).

Matching TF motifs to expressed TF genes in the excitatory neuron trajectory

To avoid correlation biases in closely-related TF motifs, we used a database of previously annotated clusters of putative binding motifs (Vierstra et al., 2020). For each motif cluster, we computed the pairwise Pearson correlation coefficients between chromVAR motif activity scores (computed from the annotated genome-wide sites of that cluster) and gene expression of all genes attributed to motifs in that cluster (Figure 3H). These PCCs were computed across the same pseudotime-pseudobulk samples that were used for CRE-gene linking. We then matched each gene to the motif cluster that exhibited the highest correlation with that gene (Figure 3F). We identified 24 dynamic motifs clusters representing 31 TFs whose gene loci are linked with at least one CRE and that exhibit high correlation coefficients of motif cluster activity and TF expression (PCC > 0.4) for downstream analysis (Figure 3F–J).

Calculation of motif synergy and correlation scores

We used the 'getAnnotationSynergy' chromVAR function (Schep et al., 2017) to compute pairwise synergy scores between motif clusters. These scores are defined as the difference in the variance of chromatin activity in CREs containing binding sites from two different motif clusters and the accessibility variance in a random sub-sample CREs which contain binding sites from only one of the motif clusters (the one with greater variance). This definition is based on the intuition that higher dynamics (variance) in accessibility in genomic loci where two TFs can potentially co-bind compared to loci where only one TF can bind is suggestive of a potential co-dependence of TFs. A positive synergy score hence corresponds to interactions where the accessibility variability explained by potential co-binding exceeds the variability explained by independent motif occurrence. In order to discriminate this notion of co-dependence from simple correlation in motif accessibility, we also computed correlation coefficients using the 'getAnnotationCorrelation' function in chromVAR. This score is defined as the correlation between the aggregate motif activity scores (deviation scores) computed from CREs that contain motifs from one but not the other motif cluster, respectively.

Fuzzy c-means: clustering and re-projection approach

For fuzzy clustering analysis, 1,267 seed cells were first selected at random from glial clusters (10% of single cells), with the number selected proportional to the cluster size. Pseudobulk datasets were sampled by combining these cells with their 50 nearest neighbors in scRNA PCA space. Next, 1,957 variably expressed genes were determined using the function 'findVariableGenes' from the R package 'URD'. A pseudobulk counts matrix was made by summing feature counts across the respective single cell members comprising each aggregate.

Fuzzy c-means clustering was performed on this pseudobulk matrix using the function "cmeans" from the R package 'e1071' with parameters c = 14 and m = 1.25, resulting in a gene-by-module "membership matrix" and a sample-by-module "centers matrix." To determine a 'fixed' or binarized module membership for downstream analyses, we defined a threshold membership score as the maximum score at which all genes were assigned to a cluster (threshold = 0.06). Gene ontology enrichments for each module were computed using the function 'enrichGO' from the R package 'clusterProfiler'. Module connectivity was computed between all module pairs using the Jaccard index, and modules were linked by applying a threshold of 0.2 of the Jaccard index of gene sharing. This threshold was chosen by applying the elbow method. To visualize the connections between modules, the centers matrix (sample-by-module) was used as the basis for dimensionality reduction with UMAP, using the R package 'umap'.

Finally, this process was repeated, sweeping the clustering parameters (c, m) and the membership threshold across a range of values; from c = 6 to c = 30, and from m = 1 to 2; to ensure that the structure of the resulting embedding was not overly sensitive to the clustering parameters.

Projecting ATAC-seq data into fuzzy clustering space and GPC projection

Pseudobulk samples of scATAC cells were generated using the same approach described above for gene activity scores. This matrix was subsetted to match the features (genes) of the RNA fuzzy clustering analysis. In the case of missing features, values were imputed using their median gene activity. To project ATAC-seq cells into the RNA fuzzy clustering embedding, we transposed the membership matrix and multiplied it with the gene activity-by-pseudobulk matrix. Finally, we used the "predict" function in R 'stats', with the fuzzy clustering UMAP model as the first argument, and the resulting transposed product matrix as the second, to determine the UMAP coordinates of ATAC pseudobulks.

To perform a projection operation, we took the matrix of samples X gene activity scores and multiplied it by the feature loadings from fuzzy C-means clustering (genes X loadings). The resulting matrix is fed into the same UMAP model used to create the original manifold. This gives a visualization of the similarity between projected points and landmarks on the manifold. To restrict this operation to GPCs, we forced the gene activity scores of other genes to be the median score. The same samples are thus projected twice, once taking into account all the genes, and again taking into account only GPCs. These two points are the basis for the arrow visualization used in Figure 6F. Finally, to provide a baseline for this transformation, we performed this operation using both random and defined control gene sets (Figure S8E).

Differential branch activity analysis

Branches were defined by grouping ATAC-seq pseudobulks projecting into the early part of the fuzzy clustering UMAP (into the Cyc cluster) according to their full-dataset cluster annotation, resulting in three branches. Differential gene activities were calculated using Wilcoxon rank sum tests to compare branch A to B and C, B to A and C, and C to A and B. Genes for each branch were ranked by their average $\log_2$ fold change in the differential test. The 50 most unique genes for each branch were visualized in a row-scaled heatmap.

Gene set enrichment analysis of GPCs was performed using the Kolmogorov-Smirnov test for GPC ranks in the differential test, relative to non-GPC ranks.

Motif enrichments for GPC TFs were derived by computing a Chi-square test for the enrichment of motifs in peaks linked to differential gene activities. To find the TF motifs that best correspond to GPC TF genes, the best-correlated TF motif activity (chromVAR) for each GPC TF across glial pseudobulks was used.

Characterization of astrocyte heterogeneity

We computed motif enrichments between peaks linked to modules 13 and module 14, which both contained *AQP4*, *APOE*, and *ALDH1* as members, using a chi-square test. Resulting *P*-values were adjusted for multiple testing using a Bonferroni correction. Next, to define groups of astrocyte cells (samples in contrast to astrocytic gene signatures (modules)), we re-clustered the RNA-derived glial pseudobulk samples, and performed unbiased differential expression testing using "DESeq2" between clusters c0 and c5, which highly expressed astrocyte genes (Zhang et al., 2016). A stringent FDR (Benjamini-Hochberg) of 1e-20 was invoked to call differential genes, since applying the DESeq2 (Love et al., 2014) framework to pseudobulks deflated *P*-values. The top 200 most differential genes were used to plot aggregate differential gene expression within the alternative dataset.

De novo *mutation filtering*

De novo mutations from 1902 children with Autism and their unaffected siblings from the Simons Simplex Collection was obtained from (An et al., 2018). From the list all mutations of coding or splice consequence as annotated by Gencode v27 (https://www.gencodegenes.org/human/release_27.html) were ignored from final analysis. Additionally, *de novo* mutation calls that are observed in gnomAD (Karczewski et al., 2020), in nonstandard chromosomes, within the low complexity repeat regions from the UCSC browser table RepeatMasker (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/database/rmsk.txt.gz) were removed from downstream analysis. Also, *de novo* mutations appearing in both affected and unaffected siblings and multiple SSC families (that is, non-singleton *de novo* mutations) were removed.

A deep learning model to predict cell-type specific chromatin accessibility from DNA sequence

BPNet is a sequence-to-profile convolutional neural network that uses one-hot-encoded DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]) as input to predict single nucleotide-resolution read count profiles. The models take in a sequence context of 2114 bp around the summit of an ATAC-seq peak and predict the per-base read probability and the aggregate counts of cluster-pseudobulk samples for the middle 1000 bp. The BPNet model also uses an input Tn5 bias track which is concatenated to the prefinal layer as explained below. The BPNet model is very similar in architecture to that employed by (Avsec et al., 2020). Briefly, our model employs a two-output task following a single channel of dilated convolutions and predicts the logits with a multinomial count loss for the per base read probability in one output and the total Tn5 insertion count for the 1000bp region with a mean squared error (MSE) loss for the other output task. The model architecture consists of 8 dilated residual convolution layers, with 500 filters in each layer. At each layer, the Keras Cropping 1D layer is used to clip out the two edges of the sequence, to match the inputs concatenated to the output of each convolution and this we naturally trim the 2114 bp sequence to a 1000 bp sequence avoiding the need for max pooling. Each dilated convolutional layer has a kernel width of 21 and the dilation increases twice for every convolutional layer starting at 1. After 8 dilated residual convolutions, to predict the total counts in the 1000 bp window the output from the last dilated convolutional layer is passed through a GlobalAveragePolling1D layers in Keras and the "tn5 bias" for the region as computed in the computational method TOBIAS (Bentsen et al., 2020) is concatenated to the output and passed through a Dense layer with 1neuron to obtain the predicted total counts. To predict the per base logits, the output from the last dilated residual convolution is append with pre-computed per base "tn5 bias" as predicted by TOBIAS and passed through a final convolution layer with a single kernel and a kernel width of 1 to predict the per base logit of reads.

This model is trained across 5 folds, each fold having different combinations of training, validation, and test chromosomes for each cluster with > 500 cells as identified by our single cell analysis. The threshold was made on the minimum number of cells to prevent the models from learning spurious binding events in very small numbers of cells. The model's performance is evaluated using two different metrics for the two output tasks separately. For the total counts predicted for the 1000 bp region, the model's performance is computed with the Spearman correlation of predicted counts to actual counts. The per-base read count track is evaluated using the Jensen-Shannon divergence distance, which computes the divergence between two probability distributions; in this case the actual per base read profile for the 1000bp region and the predicted per base read profile for the 1000bp region.

Once trained, we interpreted the model using Shap.deep_explainer (https://shap.readthedocs.io/en/latest/), which uses a modified version of the DeepLift algorithm (Shrikumar et al., 2019) to understand the features learnt by the neural network models. DeepLift computes the feature attribution of each base in an input sequence to a specific output prediction from the neural network model. In this case, the DeepLift computes the per base importance scores in the input sequence to predict the per base read count and total counts in the 1000bp separately.

GC matched training

We initially trained the BPNet models only on the cluster specific peaks, without including any negative regions. Scoring the ASD denovo mutations using these models trained only on positive peaks, we observe a strong shift in model disruption scores for the case mutation when compared to the control mutations (Cluster C10 fold 0 model rank-sum *P*-value: 6.2e-07). Given that the odds ratio of loss of function (LoF) mutations between cases and control in ASD is 2.09 (n cases = 81, n controls = 38, Fisher's exact *P*-value = 0.000147303) and majority of the *de novo* mutations are autosomal dominant, we expect a very small fraction of the mutations to be highly disrupting regulation. We observe a slight difference in number of GC within 2kb around the case mutations compared to control mutations (Ranksum *P*-value: 0.02826). To avoid the model overfitting on "GC" content in the sequence, we matched every accessible cluster specific peak with a matched GC non-peak region and retrained the cluster specific BPNet models by randomly sampling 10% of matched GC peaks added to the original peakset. Evaluating the new models using the GC matched negatives, we observe that there is no systematic shift in the disruption score observed (Cluster C10 fold 0 GC matched negative trained Ranksums Pvalue:0.273) as expected and hence we have used the models trained with GC matched negatives for all the subsequent ASD enrichment analysis.

Prioritizing ASD de novo mutations using a cell type specific neural network model

The filtered *de novo* mutations from both the affected and unaffected siblings, described in the previous section, is first overlapped with the open chromatin peak regions identified in the specific cell type. For each of the mutations overlapped, first the reference sequence centered around the mutation (2114 bp) is fed into the cell type specific neural network models across all 5 folds and the prediction of the total counts and the per base read probabilities are obtained. Next, the mutation is installed in the middle of the modeled region keeping the rest of the context sequence the same. Output predictions are obtained from the 5-fold trained models. We then compute the sum of perturbation in the per-base read count predicted by the model for the mutation for 100 bp around the mutation using the formula:

$$\sum_{k=-100}^{100} a_k - b_k,$$

where

$$a_k = exp(log\ counts\ predicted\ for\ ref.\ allele) \times softmax(read\ count\ for\ ref.\ allele\ at\ base\ k),$$

and

$$b_k = exp(log\ counts\ predicted\ for\ alt.\ allele) \times softmax(read\ count\ for\ alt.\ allele\ at\ base\ k).$$

Because we predict the log of the total counts, we first exponentiate it and multiply it by the SoftMax of the per base logits predicted by the model for the reference and the alternate sequence to compute the sum of their per base differences. This is carried out across all the 5 folds to obtain a mean score for the perturbation effect of each overlapping *de novo* mutation in the specific cluster. We prioritized mutations across all clusters with a local perturbation in counts > 30 and observed that the odds ratio for the models improves as we further increase the threshold.

Deriving the threshold for prioritization of de novo mutations in ASD cohort

The procedure involved building a null model by simulating how ablating an entire motif would affect the model predictions: (1) We found all instances of JASPAR motifs in accessible sequences (peaks). (2) For each sequence containing a motif instance, we used the model to compute a predicted accessibility counts centered around that motif. (3) For each sequence, we then ablate the motif instance in the sequence (by setting those bases overlapping the motif instance to 0) and recompute predicted accessibility counts using the model. (4) For each sequence, the difference between the two predicted counts (before and after motif instance ablation), divided by the length of motif was computed as the disruption score. We eliminated motif instances that had disruption scores less than 5 counts (these are instances with no effect on accessibility). The median of the distribution of the remaining disruption scores (30) was used as the threshold for the enrichment analysis

Calculating enrichments of motifs at predicted high effect size mutations

We overlapped all the predicted high effects case and controls mutations with JASPAR motif instances and called motif instances for mutations. To resolve ties among multiple motifs matching a mutation, we scored the motif instances overlapping a mutation with the cluster specific model that scored the mutation the highest, the per base importance scores using DeepLift normalized by the length of the mutation and picked the motif with the highest score as the disrupted motif for the mutation.

Immunohistochemistry

Immunohistochemistry was performed as described (Trevino et al., 2020). Briefly, PCW17 and PCW21 human cortical tissue was fixed overnight at 4°C in 4% paraformaldehyde (PFA, Electron Microscopy Sciences). Samples were then washed with PBS and transferred to a 30% sucrose solution for 48-72 h, then embedded in OCT (Tissue-Tek OCT Compound, 4583, Sakura Fenetek) and 30% sucrose at a 1:1 ration, and snap-frozen in dry ice. Cryosections were obtained using a cryostat (Leica) set at 30 μm and mounted on Superfrost Plus Micro slides (VWR, 48311-703). Next, sections were blocked and permeabilized for 1 h at room temperature in blocking solution (10% normal donkey serum, 0.3% Triton-X in PBS) and incubated with primary antibodies diluted in the same solution overnight at 4°C. The following primary antibodies were used: anti-ASCL1(Mouse, 1:100, BD Biosciences, 556604), anti-CTIP2 (Rat, 1:300, Abcam, ab18465), anti-EGFR (Rat, 1:200, Abcam, ab231), anti-GFAP (Rabbit, 1:1,000, Dako, Z0334), anti-GFAP (Rat, 1:1000, Thermo Fisher Scientific, 13-0300), anti-HOPX (Mouse, 1/50, Santa Cruz, sc-398703), anti-KI67 (Mouse, 1:500, BD Biosciences, 550609), anti-OLIG2 (Rabbit, 1:200, Millipore, AB9610), anti-PBXIP1 (Rabbit, 1:100, Abcam, ab84752), anti-PDGFRA (Rabbit, 1:200, Santa Cruz, sc-338), anti-PPP1R17 (Rabbit, 1:200, Atlas Antibodies, HPA047819), anti-SOX9 (Goat, 1:500, R&D Systems, AF3075), anti-SPARCL1 (Goat, 1:300, Novus Biologicals, AF2728), anti-TFAP2C (Rabbit, 1:100, Thermo Fisher Scientific, 14572-1). PBS was used to wash off the primary antibodies, and sections were then incubated with Alexa Fluor secondary antibodies (1:1,000, Life Technologies) for 1 h at room temperature. Hoechst 33258 was used to visualize the nuclei. Sections were mounted for microscopy with glass coverslips using Aquamount (Thermo Scientific). Images were taken using a Leica TCS SP8 confocal microscope and processed using ImageJ (Fiji). Cortical images spanning from VZ to CP were obtained using a tiling approach in the Leica TCS SP8 and automatically stitched using the Leica software.

QUANTIFICATION AND STATISTICAL ANALYSIS

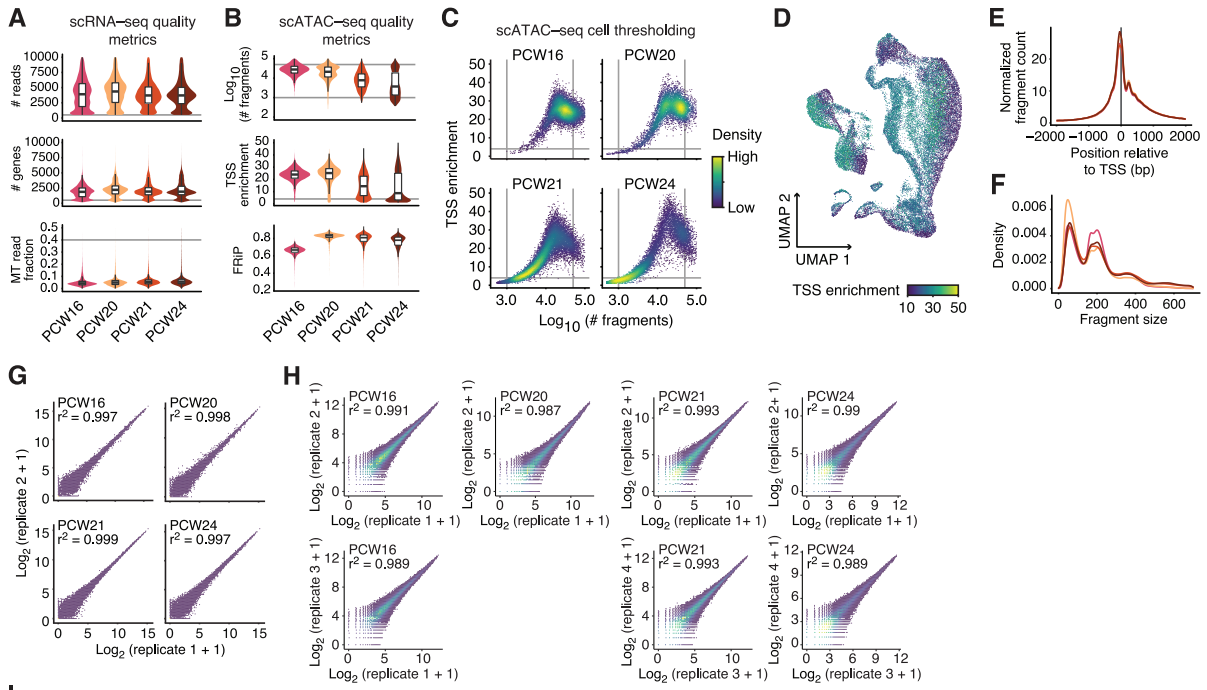
All statistical analyses were performed in R v3.6.3 or Python v3.8. Statistical tests are described in the relevant methods sections above. Significant cell-type specific enrichment for prioritized *de novo* mutations were determined using a Fisher's Exact Test with Bonferroni correction.

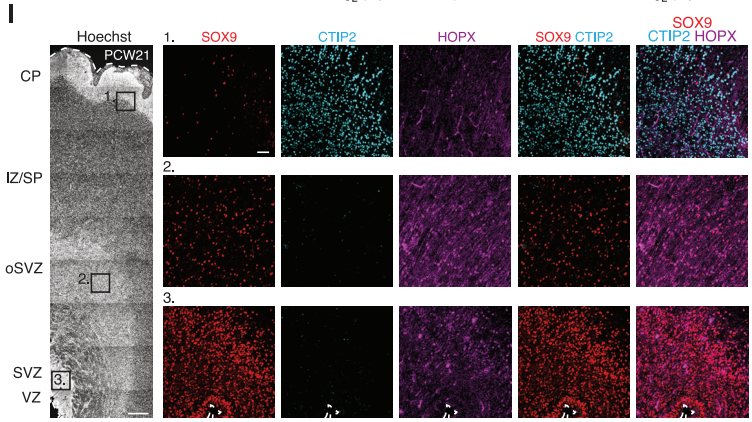
ADDITIONAL RESOURCES

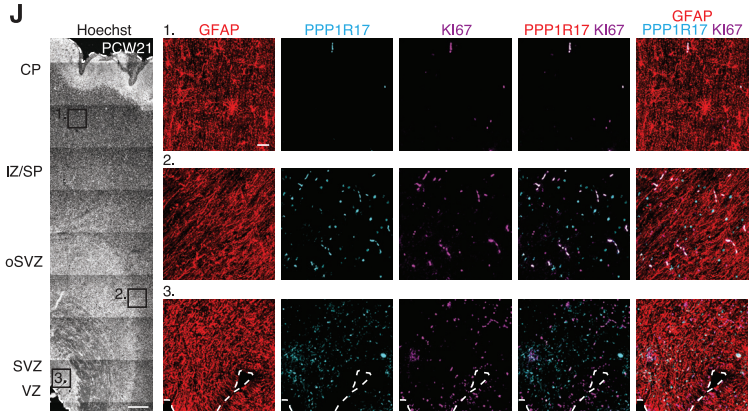
https://scbrainregulation.su.domains

CellPress

Supplemental figures







(legend on next page)

Figure S1. Data quality of scATAC-seq and scRNA-seq libraries and immunohistochemistry of human cerebral cortex architecture, related to Figure 1

(A) scRNA-seq quality metrics showing the distribution of the number of reads, number of genes, and mitochondrial (MT) gene fraction per cell in each sample. Technical replicates are merged. PCW = postconceptional weeks.

(B) scATAC-seq quality metrics showing the distribution of the number of fragments, transcription start site (TSS) enrichment, and fraction of reads in peaks (FRiP) per cell in each sample.

(C) scATAC-seq cell thresholding on TSS enrichment and fragment counts.

(D) UMAP plot showing the TSS enrichment of each cell.

(E) Aggregate normalized fragment count around TSSs for each scATAC-seq sample.

(F) Aggregate fragment size distributions for each scATAC-seq sample.

(G) Correlation of technical replicates for each scRNA-seq sample.

(H) Correlation of technical replicates for each scATAC-seq sample.

(I) Immunohistochemistry in PCW21 human fetal cerebral cortex, showing expression of SOX9, CTIP2, and HOPX in the ventricular zone (VZ), subventricular zone (SVZ), outer SVZ (oSVZ), intermediate zone / subplate (IZ/SP), and cortical plate (CP). This image was generated by automatic stitching of individual images.

(J) Immunohistochemistry in PCW21 human fetal cerebral cortex, showing expression of GFAP, PPP1R17, and KI67. This image was generated by automatic stitching of individual images.

Scale bars, 500 μm (A, B), 50 μm (insets A, B).

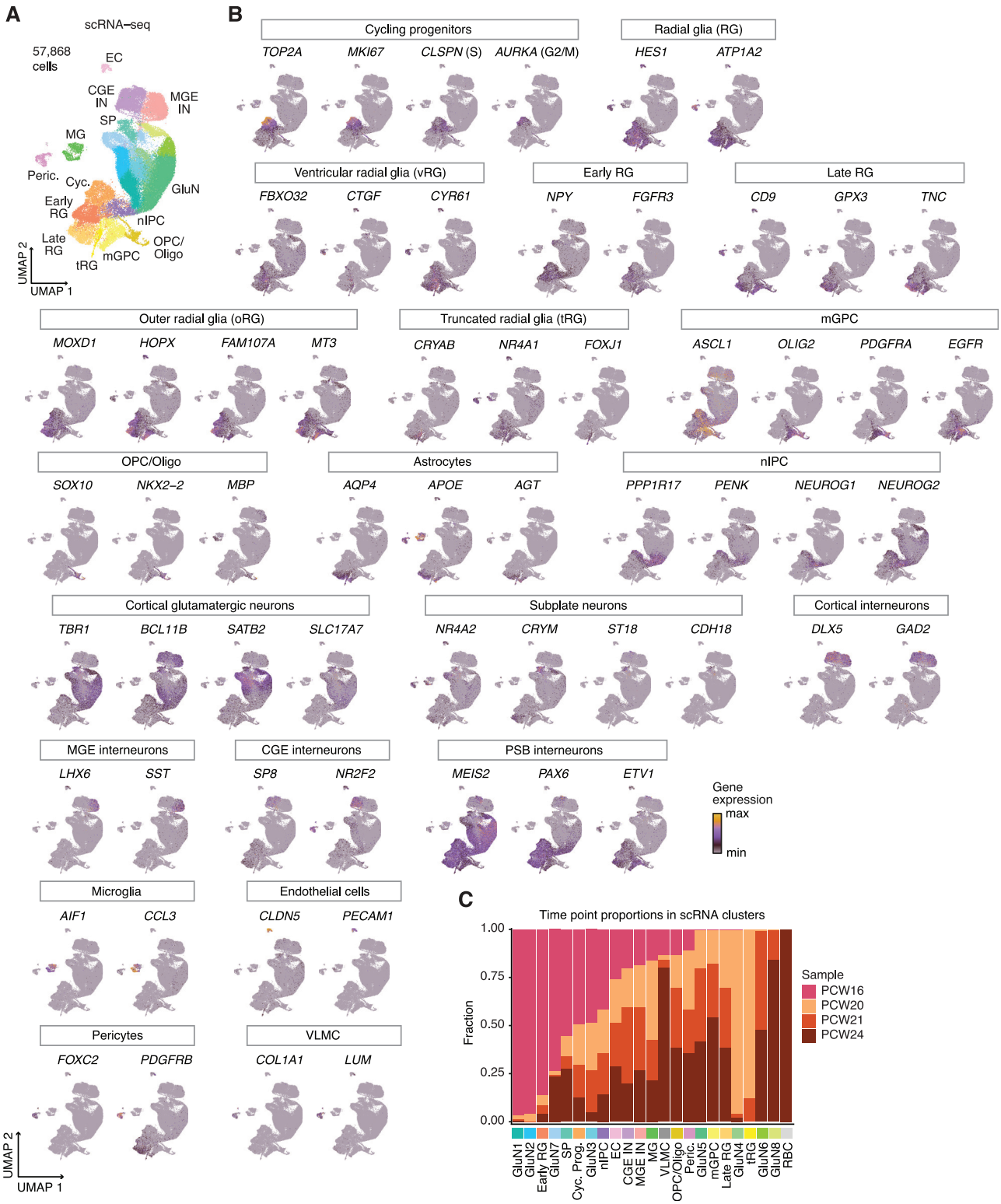
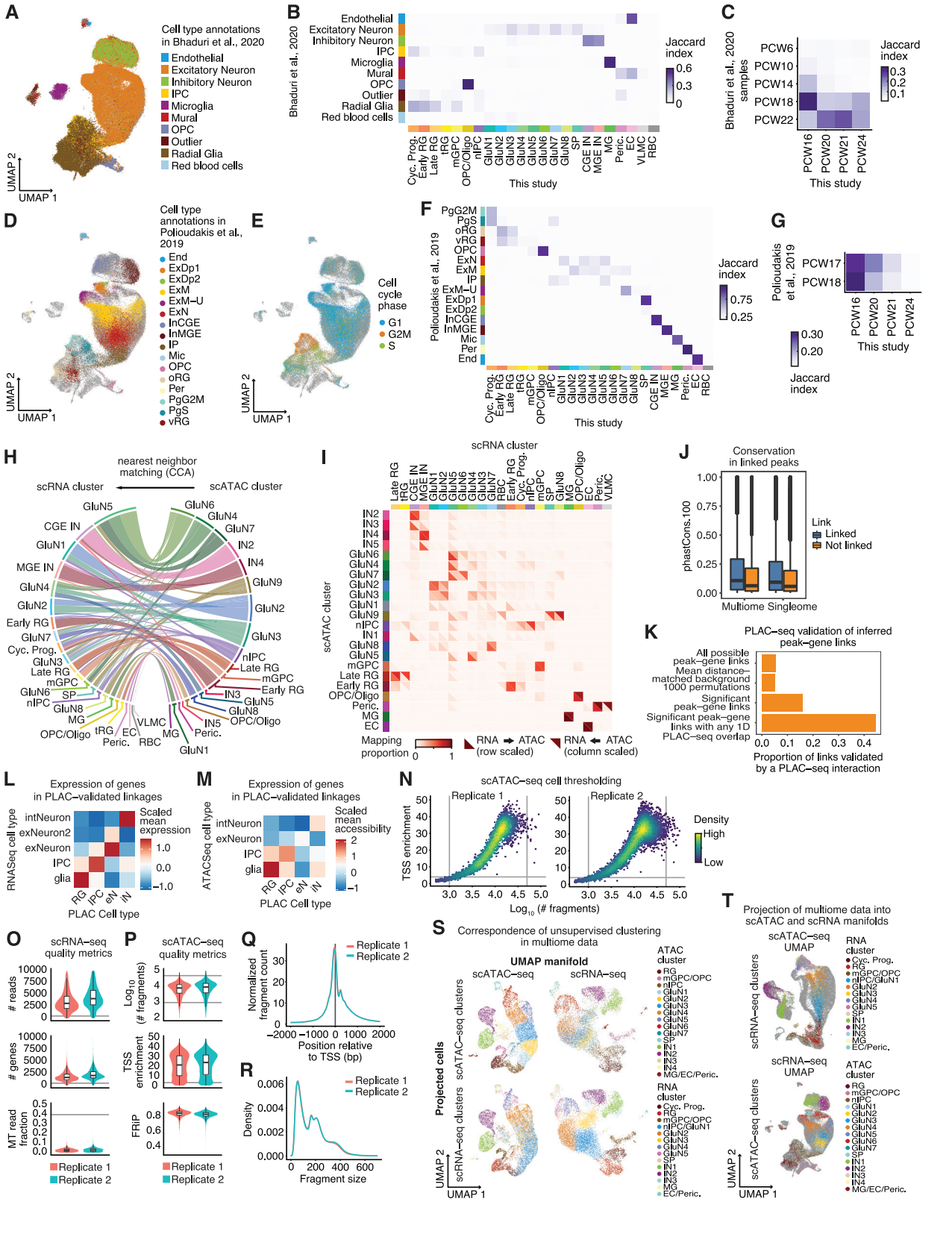


Figure S2. Expression of cell-type-specific markers in scRNA-seq data, related to Figure 1

(A) UMAP of scRNA cells colored by cluster (from Figure 1F).

(B) UMAP plots showing gene expression of cell-type and cluster-specific markers.

(C) Bar plot showing the sample age composition in each of the scRNA-seq clusters.



(legend on next page)

Figure S3. Comparison with external scRNA-seq dataset from human cortex, linkage of scRNA-seq and scATAC-seq datasets, evaluation of peak-to-gene links and data quality of scATAC-seq and scRNA-seq multiome data, related to Figure 2

(A) Projection of alternate data from Bhaduri et al., 2020 into this scRNA-seq manifold, showing alignment of broad cell types.

(B) Jaccard index of genes expressed in clusters from this scRNA-seq dataset and annotated cell types from Bhaduri et al. (2020).

(C) Correspondence of gestational age to projected annotation (Bhaduri et al., 2020). Annotations were matched using nearest-cell matching in projected space (Methods).

(D) Projection of neocortical fetal cells (Polioudakis et al., 2019) into scRNA-seq manifold. Colors show annotated cell types.

(E) Projection of neocortical fetal cells (Polioudakis et al., 2019) into scRNA-seq manifold. Colors show annotated cell cycle phase.

(F) Correspondence of cluster annotation and cell type annotation of projected data. Annotation was matched using nearest-cell matching in projected space (Methods).

(G) Correspondence of gestational age to projected annotation (Polioudakis et al., 2019). Annotations were matched using nearest-cell matching in projected space (Methods).

(H) Ribbon plot showing correspondence of scRNA-seq and scATAC-seq clusters in a shared canonical correlation analysis (CCA) landscape. CCA was derived from expression values in scRNA-seq data matched to gene activity scores from scATAC-seq.

(I) Confusion matrix showing the correspondence of cluster annotations across datasets in the CCA. Upper triangles indicate how ATAC clusters match to RNA clusters; lower triangles indicate how RNA clusters match to ATAC clusters. Coloring indicates the proportion of cells mapping for a given pair.

(J) PhastCons 100-way vertebrate conservation scores for gene-linked and unlinked CREs.

(K) Bar plot showing the proportion of peak-gene links validated at both anchors by interactions from a proximity ligation assisted ChIP and sequencing (PLAC-seq) dataset targeting H3K4me3. Categories included were the space of all possible peak-gene links; 1000 permutations drawn from all possible peak-gene links, where for each permutation, 10,000 peaks were selected to match the length distribution of significant links; significant peak-gene links; and significant peak-gene links with a 1-dimensional overlap with any PLAC-seq region. Relative to all possible links, Fisher's exact test OR = 2.91, $p < 2.2x10^{-16}$ for any linked peaks, OR = 8.09, $p < 2.2x^{-16}$ for linked peaks represented in PLAC-seq data)

(L) Heatmap showing the scaled average expression of linked genes in the major cell types in this scRNA-seq dataset, partitioned by the sorted PLAC-seq cell types in which the linkage was validated. For example, expression of the genes in PLAC-seq-validated peak-gene links, where the PLAC-seq interaction was called in radial glia, are shown in the first column. RG, radial glia; IPC, intermediate progenitor cell; eN, excitatory neuron; iN, inhibitory neuron. For row names, exNeuron, excitatory neurons, comprising GluN clusters; inNeuron, inhibitory neurons.

(M) As in (**L**), with heatmap values indicating the scaled average accessibility of peaks in PLAC-seq validated peak-gene links.

(N) scATAC-seq cell thresholding on TSS enrichment and fragment counts.

(O) scRNA-seq quality metrics showing the distribution of the number of reads, number of genes, and mitochondrial (MT) gene fraction per cell in each biological replicate. Technical replicates are merged.

(P) scATAC-seq quality metrics showing the distribution of the number of fragments, transcription start site (TSS) enrichment, and fraction of reads in peaks (FRiP) per cell in each biological replicate.

(Q) Aggregate normalized fragment count around TSSs for each scATAC-seq biological replicate.

(R) Aggregate fragment size distributions for each scATAC-seq biological replicate.

(S) UMAP embeddings for multiome scATAC (left panels) and multiome scRNA (right panels). Cells are colored by unsupervised clustering of scATAC counts (top panels) and scRNA data (bottom panels).

(T) Projection of multiome scATAC and scRNA data into singleome scATAC (top) and scRNA (bottom) UMAP manifolds.

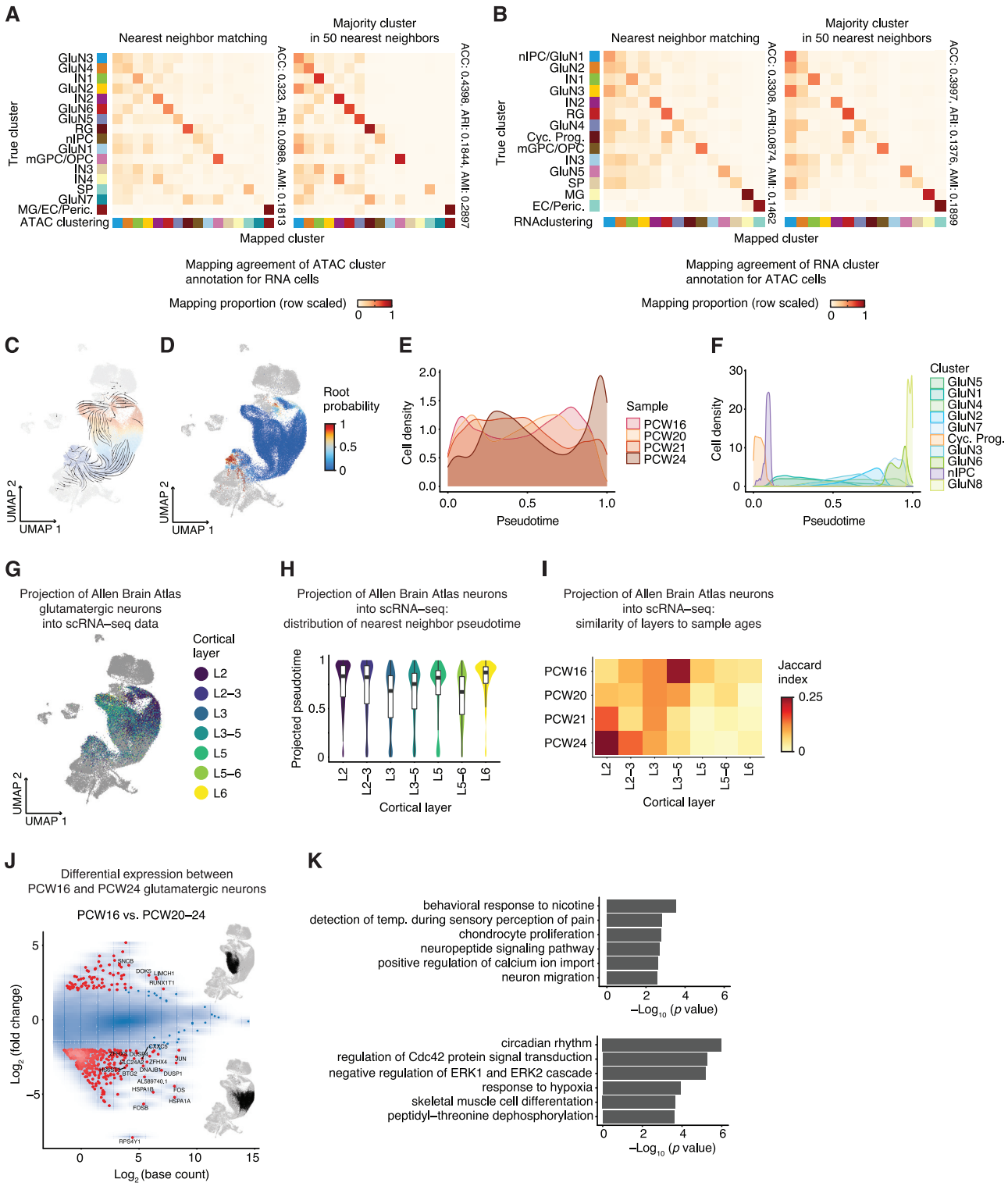


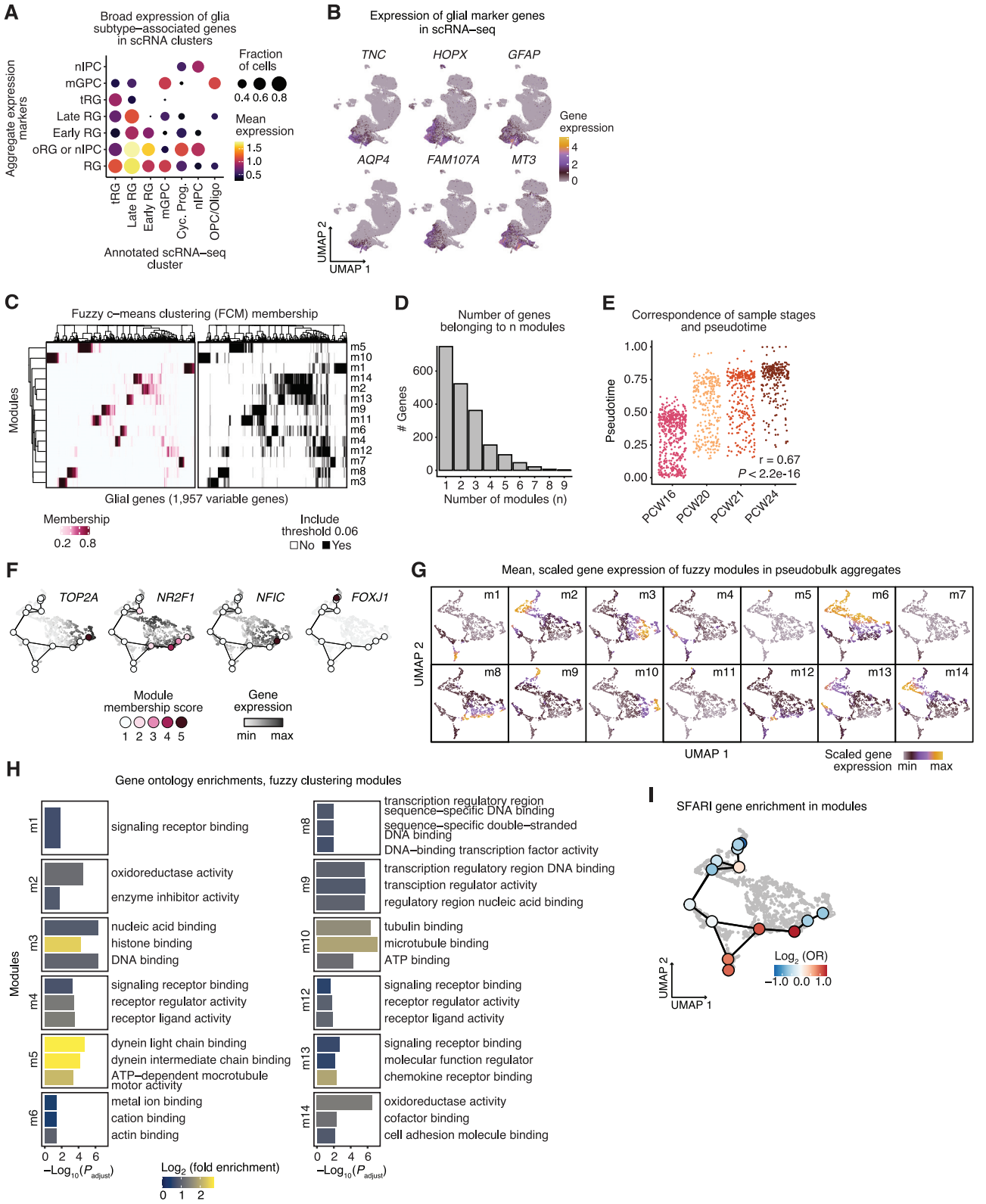
Figure S4. Supplemental analyses to glutamatergic neuron developmental trajectories, related to Figure 3

(A) Cluster agreements between true ATAC clusters and CCA-mapped clusters in multiome data, computed by matching each cell's RNA-derived gene expression profiles with its nearest (left) or 50 nearest (right) ATAC-derived gene activity profiles in CCA space. Heatmap shows the proportion of cells in each true cluster (rows) corresponding to mapped clusters (columns). Accuracy (ACC), adjusted Rand Index (ARI), and adjusted Mutual Information (AMI) are shown.

(B) Cluster agreements as above between true RNA-derived clusters and CCA-mapped RNA clusters in multiome data.

(legend continued on next page)

(C) RNA velocity streamplot in UMAP space. Aggregate velocities for cells in clusters for excitatory neuron trajectories were computed and plotted using scVelo.
(D) scVelo root probability in UMAP space. Probabilities were computed using the scVelo function 'scv.tl.terminal_states'.
(E) Density plot of sample age for individual cells along the excitatory neuron trajectory pseudotime.
(F) Density plot of cell clusters along the excitatory neuron trajectory pseudotime.
(G) Projection of adult glutamatergic neurons into scRNA UMAP space. Adult scRNA data and cortical layer annotation for cells was obtained from the Allen Brain Atlas.
(H) Distribution of excitatory neuron trajectory pseudotime for annotated cortical layers. Fetal cell pseudotime annotation was transferred to adult neurons by nearest-neighbor matching in UMAP space.
(I) Correspondence between fetal sample age and annotated adult cortical layers. The heatmap shows Jaccard indices of annotation in adult neurons with fetal gestational age annotation by nearest neighbor matching in UMAP space
(J) MA plot of differential expression between PCW16 and PCW20-24 cells. Densities for differential expression statistics for each gene are shown. Genes identified as differentially expressed are shown in red (adjusted p value < 0.05, $|\log2(\text{fold-change})| > 2$). Cells with $0.2 \leq$ annotated pseudotime ≤ 0.8 were compared in PCW16 versus PCW20, PCW21 and PCW24. Differential expression was computed by applying DESeq2 on 20 randomly sampled pseudobulk samples for each age group (100 cells each).
(K) GO enrichment for genes upregulated (top) and downregulated (bottom) in PCW16 versus PCW20-24 neurons. Enrichments were computed for the gene sets shown in H and the top 6 enrichments are shown for each direction.



(legend on next page)

Figure S5. Glial cell characterization using fuzzy c-means clustering, related to Figure 4

(A) Bubble plot showing gene expression of glial subtype markers in annotated glial clusters. The expression of identifying markers is sometimes evident in several clusters. For each group of markers, the dot size indicates the mean fraction of cells expressing the markers. Color indicates mean expression level.

(B) UMAP showing expression of selected glial genes in the scRNA-seq manifold.

(C) Membership matrix for fuzzy clustering, showing the fractional membership of each gene (columns) in each module (rows). The right-hand panel shows the memberships, now binarized at a membership threshold of 0.06.

(D) Bar plot showing how many genes belong to "n" modules after thresholding.

(E) Plot of glial scRNA-seq pseudobulk aggregates. For each aggregate, the sample-of-origin age in postconceptional weeks (PCW) is compared with the pseudotime values (Methods). Pseudotime was strongly correlated with developmental time. Pearson $r = 0.67$, $p < 2.2x10^{-16}$.

(F) Module membership and expression values for genes depicted in Figure 4C across pseudotime aggregates.

(G) UMAP plots showing the mean, scaled expression of all genes in each module (m1-m14).

(H) Gene ontology (GO) enrichments for each module, including the term description. Bar plots represent the $-Log_{10}(P)$, with P values adjusted by the Bonferroni method. Bar color indicates the log_2 fold enrichment for each term. Modules 7 and 11 exhibited no significant enrichments at $q < 0.05$.

(I) Enrichment of SFARI genes (gene score < 3) in each fuzzy module. Enrichments indicated by color, are shown as the log_2 odds ratio (OR), and plotted with module centroids in the UMAP of fuzzy clustering cell loadings.

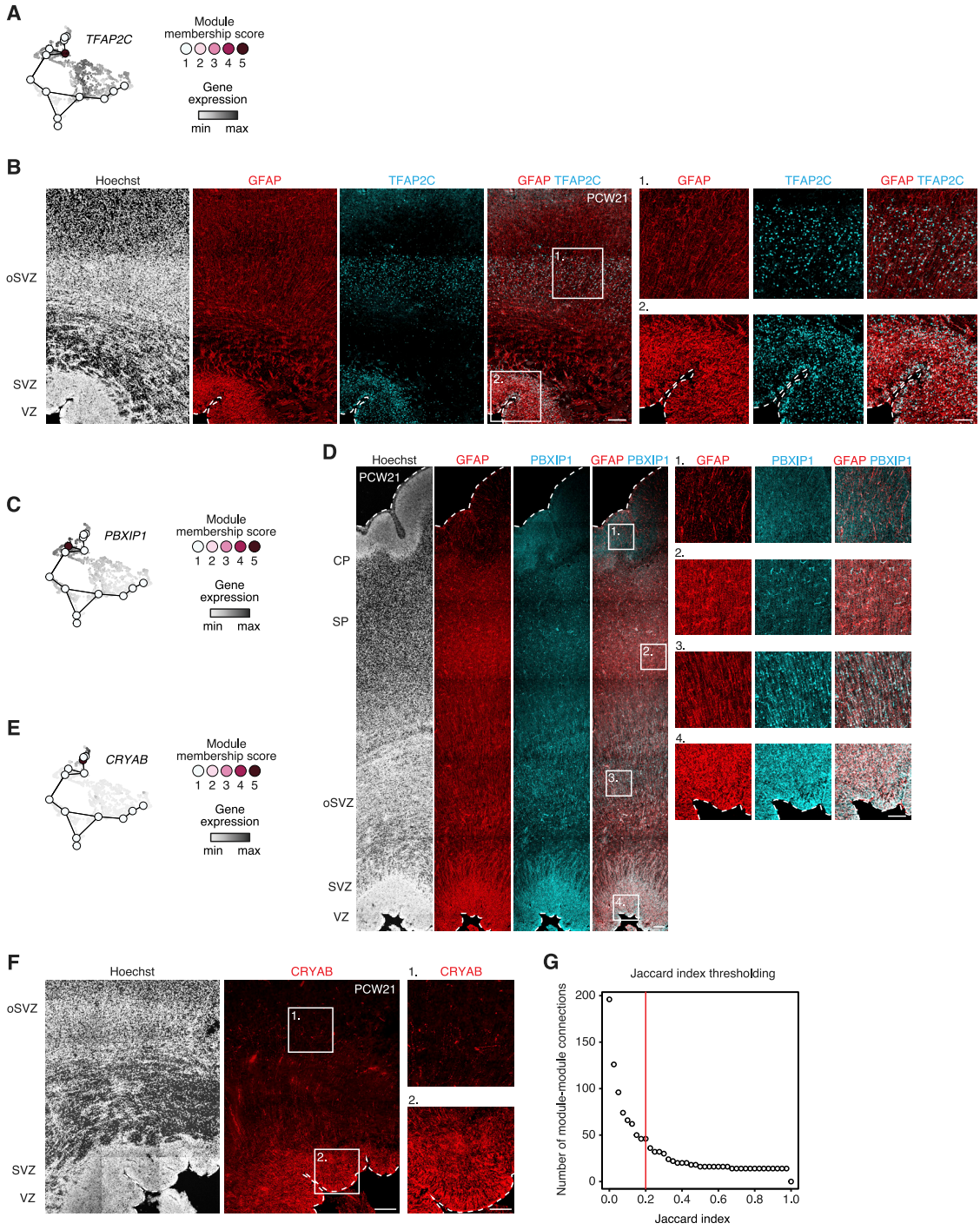


Figure S6. Immunohistochemistry of genes in fuzzy modules, Related to figure 4

(A) Module membership and expression values for *TFAP2C*.

(B) Immunohistochemistry in PCW21 human cerebral cortex showing expression of module m6 transcription factor TFAP2C in the SVZ and oSVZ. This image was generated by automatic stitching of individual images.

(C) Module membership and expression values for *PBXIP1*.

(D) Immunohistochemistry in PCW21 human cerebral cortex showing expression of module m2 marker PBXIP1 and colocalization with the astroglia marker GFAP in radial glia in the VZ and oSVZ. This image was generated by automatic stitching of individual images.

(E) Module membership and expression values for *CRYAB*.

(legend continued on next page)

(F) Immunohistocjemistry in PCW21 human cerebral cortex showing expression of module m9 marker CRYAB in truncated radial glia in the VZ. This image was generated by automatic stitching of individual images.
(G) Plot of the total number of module-module connections at a given Jaccard index threshold. Higher Jaccard thresholds mean fewer connections are "allowed" in the downstream analysis. This plot shows a clear "elbow" behavior at Jaccard > 0.2, which was used to select that threshold.
Scale bars, 100 μm (insets B, D, F), 200 μm (B, D, F).

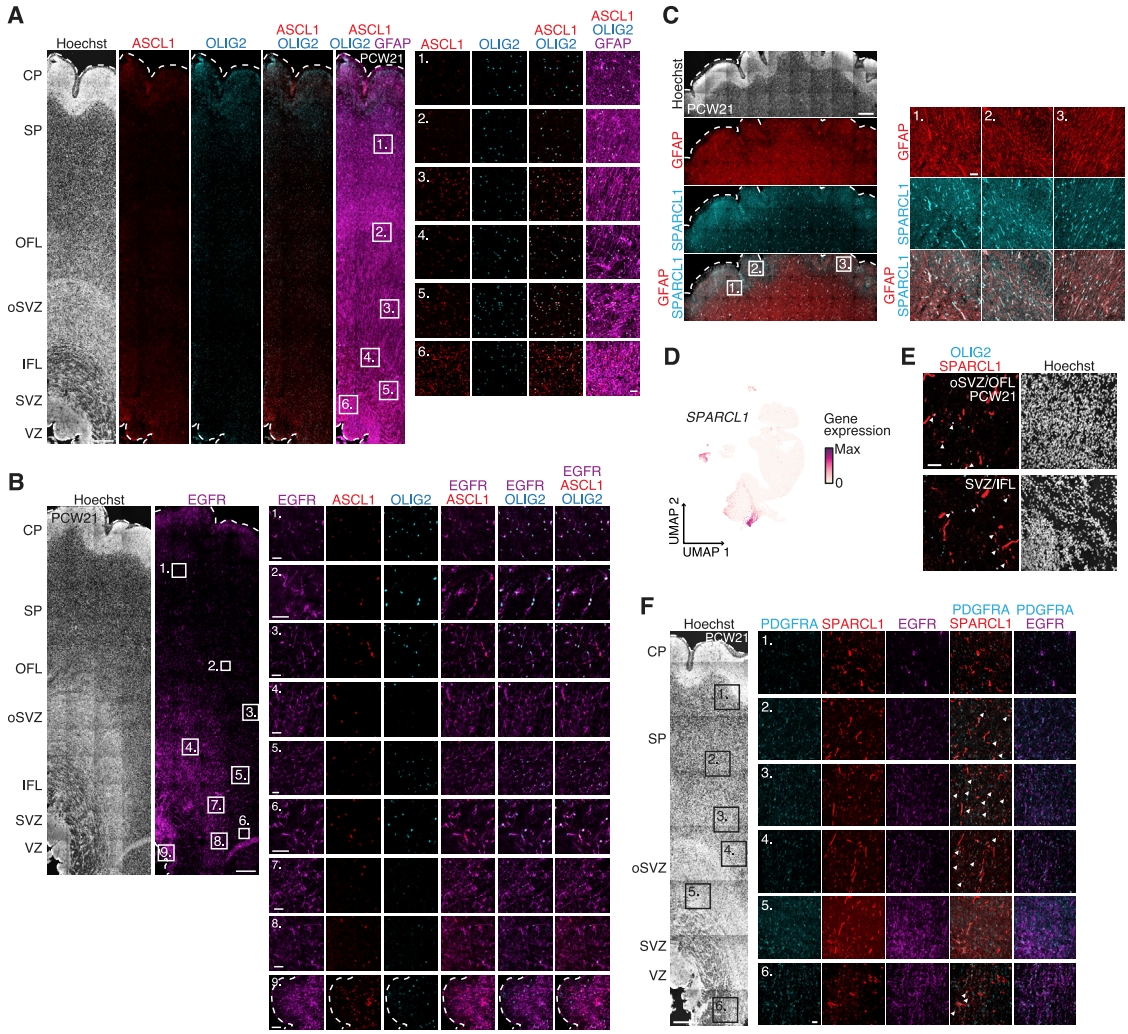


Figure S7. Characterization of mGPC and colocalization of astrocyte- and oligodendrocyte-associated markers in the human cerebral cortex, related to figure 4

(A) Immunohistochemistry in PCW21 human fetal cortex showing expression of ASCL1, OLIG2 and GFAP. ASCL1 and OLIG2 colocalize in the inner and outer fiber layers (IFL, OFL) and SVZ and oSVZ mainly. GFAP shows the radial glial scaffolding. This image was generated by automatic stitching of individual images.

(B) Immunohistochemistry in PCW21 human fetal cortex showing expression and colocalization of modules m1, m4 and m12 genes ASCL1, OLIG2 and EGFR representing mGPCs. This image was generated by automatic stitching of individual images.

(C) Immunohistochemistry in PCW21 human fetal cortex showing colocalization of the astroglia markers SPARCL1 and GFAP in the cortical plate and subplate. This image was generated by automatic stitching of individual images.

(D) UMAP plot showing *SPARCL1* gene expression.

(E) Immunohistochemistry in PCW21 human fetal cortex showing colocalization (white arrowheads) of OLIG2, associated with oligodendrocyte progenitors, and the astrocyte marker SPARCL1 in SVZ/IFL and oSVZ/OFL.

(F) Immunohistochemistry in PCW21 human fetal cortex showing colocalization of PDGFRA, SPARCL1 and EGFR. This image was generated by automatic stitching of individual images.

Scale bar, 50 μm (E, insets A, B, C, F), 500 μm (A, B, C, F).

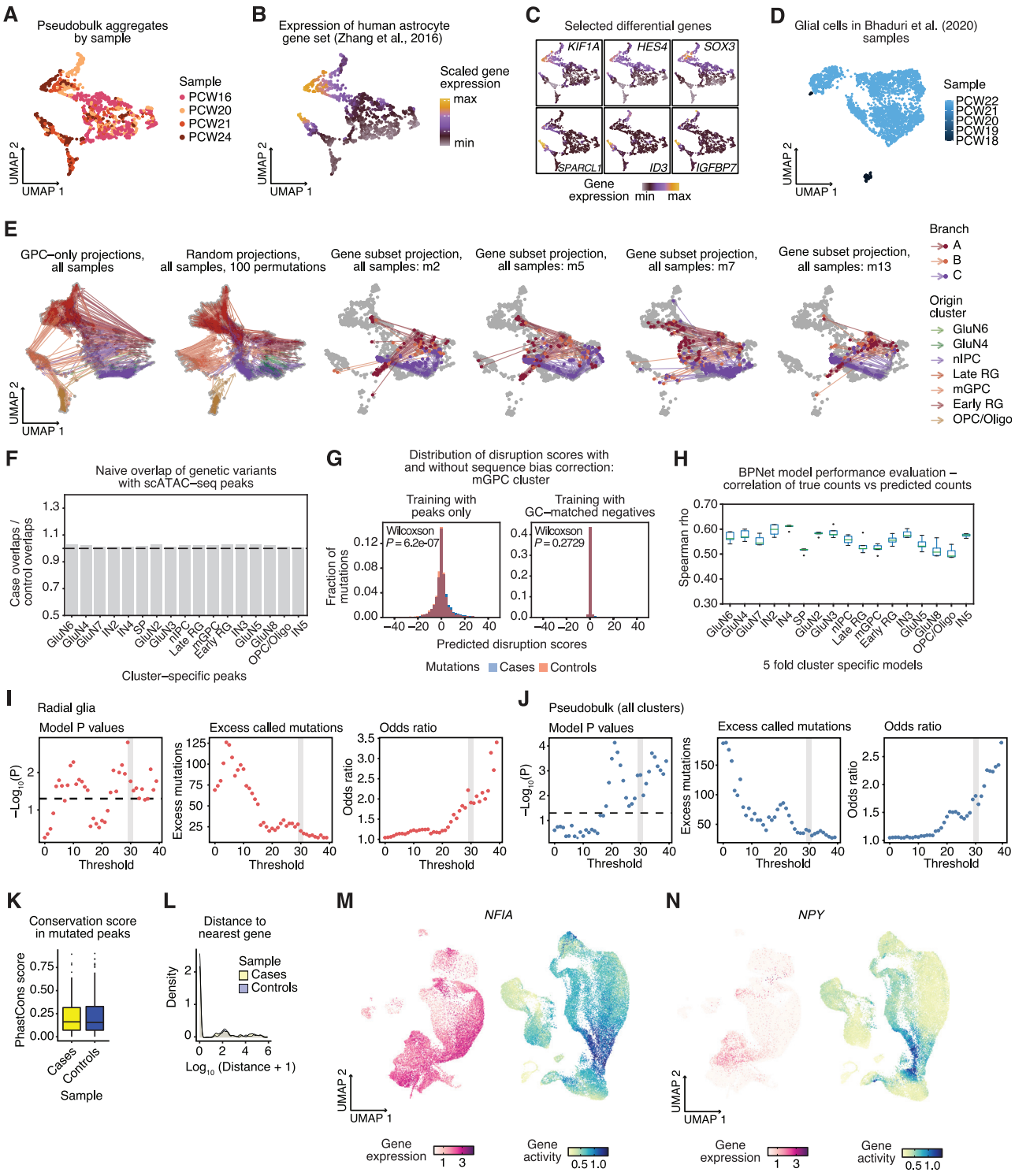


Figure S8. Heterogeneity of astrocyte precursors, projection of scATAC-seq aggregates into fuzzy embedding, and characterization of BPNet model performance and mutation vignettes, related to Figures 5–7

(A) Fuzzy clustering-derived UMAP showing pseudobulk aggregates plotted by sample age.

(B) Mean scaled expression of human mature astrocyte genes (Zhang et al., 2016) in fuzzy clustering-derived UMAP of scRNA-seq pseudobulk aggregates.

(C) Expression of selected differential genes from Figure 5D.

(D) UMAP of Bhaduri et al., 2020 fetal astrocyte scRNA-seq dataset, showing sample age.

(legend continued on next page)

(E) UMAP plots showing the projection of aggregates into the fuzzy clustering-derived low-dimensional embedding. The origin of the arrows represents the original projection coordinates of a particular scATAC-seq aggregate; the arrows point to the new projection coordinates when using only a given subset of genes to make the projection (other genes are imputed as zero-variance features). Colors indicate the scATAC-seq cluster from which the aggregates derive. This panel shows projection with only GPC genes (Methods). Subsequent panels show projections with: Random gene sets (100 permuted trials); module m2 genes only; module m5 genes; Module m7 genes; module m13 genes.
(F) Enrichment of cases versus control mutations using naive overlap with cluster-specific ATAC-seq peaks, showing relevance of the deep learning model to capture pathogenic disruptions.
(G) Distribution of disruption scores for case and control mutations using different training paradigms. Data are shown for the mGC cluster. On the left, using only scATAC-seq peaks as the basis for training, there is a systematic difference between cases and controls (Wilcoxson test $p = 6.2x10^{-7}$). On the right, when training is given GC-matched negatives, disruption scores are substantially more conservative, and the distributions are matched ($p = 0.27$).
(H) Performance evaluation of BPNet cluster-specific models, computed by calculating the rank correlation between true counts in the cluster and predicted counts. Data are from 5-fold cross-validated training.
(I) Evaluation of robustness in disease prioritization of Radial Glia model across different threshold values. From left to right, the -log(Fisher's exact test p value), excess in number of causal mutations observed in cases compared to controls and the Fisher's exact test odds ratio are plotted across all threshold values.
(J) Evaluation of robustness in disease prioritization of the pseudobulk model across different model threshold values. From left to right, the -log(Fisher's exact test p values), excess in number of causal mutations observed in cases compared to controls and the Fisher's exact test odds ratio are plotted across all threshold values.
(K) Conservation scores in cases versus controls, showing that trivial genomics metrics do not explain the observed prioritized mutations.
(L) Distance to the nearest gene in cases versus controls, showing that trivial genomics metrics do not explain the observed prioritized mutations.
(M) UMAP plots of gene expression (magenta) and gene activity scores (viridis) for *NFIA*.
(N) UMAP plots of gene expression (magenta) and gene activity scores (viridis) for *NPY*.