nature biotechnology

Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain

Blue B Lake[1,6], Song Chen[1,6], Brandon C Sos[1,2,6], Jean Fan[3,6], Gwendolyn E Kaeser[2,4], Yun C Yung[4], Thu E Duong[1,5], Derek Gao[1], Jerold Chun[4], Peter V Kharchenko[3] & Kun Zhang[1]

Detailed characterization of the cell types in the human brain requires scalable experimental approaches to examine multiple aspects of the molecular state of individual cells, as well as computational integration of the data to produce unified cell-state annotations. Here we report improved high-throughput methods for single-nucleus droplet-based sequencing (snDrop-seq) and single-cell transposome hypersensitive site sequencing (scTHS-seq). We used each method to acquire nuclear transcriptomic and DNA accessibility maps for >60,000 single cells from human adult visual cortex, frontal cortex, and cerebellum. Integration of these data revealed regulatory elements and transcription factors that underlie cell-type distinctions, providing a basis for the study of complex processes in the brain, such as genetic programs that coordinate adult remyelination. We also mapped disease-associated risk variants to specific cellular populations, which provided insights into normal and pathogenic cellular processes in the human brain. This integrative multi-omics approach permits more detailed single-cell interrogation of complex organs and tissues.

The human brain is an enormously complex network comprising ~100 billion spatially organized and functionally interconnected neurons embedded in an even larger population of glia and non-neural cells. The creation of a complete cell atlas of the human brain will require highly scalable and unbiased single-cell approaches constrained by neither the availability of fresh biopsies nor the dissociation methods required to isolate living whole cells. Cell nuclei isolates provide a viable alternative, as they can be derived from fresh or archived tissues, provide sufficient RNA for accurate prediction of cellular expression levels[1–4], and are free of artifacts associated with tissue dissociation[5]. We recently demonstrated that single-nucleus transcriptome sequencing (SNS) can resolve neuronal subtype diversity across multiple human cortical brain regions[4] at a relatively high sequencing depth (~8 million reads per cell). However, scaling-up was limited by throughput (maximally 96 cells per microfluidic chip), high cost, and sampling bias arising from poor capture of smaller non-neuronal nuclei on microfluidic chips. Our study showed that higher-throughput single-nucleus RNA-seq approaches specifically applicable to archived human tissues are needed.

Although transcriptomic profiling allows the identification of functionally distinct cell types that make up complex tissues, the inclusion of epigenetic information can provide a more complete picture of how these expression profiles are regulated or maintained. Genome-wide studies have mapped regulatory sites to open or hyperaccessible chromatin within gene promoter and enhancer regions, revealing shared *cis*-regulatory sites that can distinguish cell types and lineages[6,7]. The identification of such cell-type-specific regulomes will lead to improved understanding of the genetic programs that define cellular differentiation, commitment, and functionality. Furthermore, because common genetic variants associated with diverse traits and diseases fall mostly within intronic or intergenic regions[8], with enrichment in tissue-specific regulatory sites[6,7], the generation of cell-type-specific regulome maps could provide additional valuable insights into the underlying mechanisms of disease. As with transcriptomic studies, a major limitation of available epigenetic assays has been the requirement for large numbers of cells. Recent methods have improved sensitivity enough to reduce this requirement to hundreds of cells[9] and sometimes even to the single-cell level[10–13]; however, the application of such single-cell methods has yet to be demonstrated at a large scale on highly heterogeneous archived human tissues such as the brain.

Ultimately, the comprehensive mapping of human brain cell types and their overall phenotypic potential necessitates more efficient methods for nuclear RNA sequencing and co-profiling of epigenomic attributes using archived tissues. As nuclear isolates are quite amenable to single-cell genomic studies[14,15], we have developed two parallel high-throughput methods for quantifying nuclear transcripts and measuring DNA accessibility at the single-cell level that are applicable to the same pool of nuclei. This provides a means for integrative analysis of gene expression and regulation in the same archived human tissue. Here we resolved extensive cellular diversity in defined regions of the human cortex and cerebellum, identified region-specific neuronal and non-neuronal cell types, and identified defining transcription factor activities and target gene expression profiles on a large scale. Finally, by mapping disease risk variants to

[1]Department of Bioengineering, University of California San Diego, La Jolla, California, USA. [2]UC San Diego School of Medicine, La Jolla, California, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. [4]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California, USA. [5]Department of Pediatric Respiratory Medicine, University of California San Diego, La Jolla, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.C. (jchun@sbpdiscovery.org), P.V.K. (Peter_Kharchenko@hms.harvard.edu) or K.Z. (kzhang@bioeng.ucsd.edu).

Received 14 April; accepted 15 November; published online 11 December 2017; doi:10.1038/nbt.4038

cell-type-specific regulatory regions, we obtained proof-of-concept identification of possible pathogenic cell types underlying several brain-related diseases.

RESULTS

Single-cell interrogation of human cortex and cerebellum

Recent advances in droplet-based technologies have greatly enhanced the throughput of single-cell RNA-seq[16–18], enabling simultaneous transcriptomic profiling of tens of thousands of single cells. Although these methodologies reduce the depth of coverage, they enable extensive classification of cell type and state, providing unique expression signatures to resolve functional heterogeneity within tissues. We adapted a droplet-based methodology[17] to analyze single nuclei, termed snDrop-seq (**Fig. 1a**, **Supplementary Fig. 1**), to permit larger-scale assessment of gene expression dynamics in live and archived human tissue. Our method specifically addressed the challenge of disrupting nuclear membranes in microdroplets without causing excessive RNA degradation. We applied snDrop-seq to adult human postmortem brain samples encompassing the visual cortex (Brodmann area 17 (BA17) or visual area 1 (V1)), the frontal cortex (BA10 and BA6), and the lateral cerebellar hemisphere from six different individuals (**Supplementary Table 1**).

To coinvestigate epigenetic configurations, we developed a single-cell DNA-accessibility assay that combined our previously developed transposome hypersensitive site–sequencing assay[9] with combinatorial cellular indexing[11] using customized barcoded transposomes (**Fig. 1a**, **Supplementary Fig. 2**). scTHS-seq takes advantage of linear amplification by *in vitro* transcription and an engineered super-mutant of Tn5 transposase[19] to achieve higher sensitivity than ATAC-seq[20], including better coverage of distal enhancers found to be highly cell-type specific[9,21]. Applying both methodologies, we profiled expression and regulation signatures from the same brain regions, which allowed independent and unbiased discovery of cellular diversity and, through integrative analyses, gene expression and regulation profiles distinctive to specialized cells (**Fig. 1a**).

We generated 36,166 single-nucleus expression measurements after quality-filtering, and we resolved 35,289 of these from the visual (19,368 nuclei) and frontal (10,319 nuclei) cortices and the cerebellar hemisphere (5,602 nuclei) into neuronal or non-neuronal cell types (**Fig. 1b**, **Supplementary Fig. 3**, **Supplementary Table 2**). Analysis of cross-species mixing confirmed a low percentage of doublets, similar to that found for whole-cell measurements (2–11%; **Supplementary Fig. 1**)[17]. We sequenced these libraries to an average of 6,200 usable reads per nucleus (**Supplementary Table 1**), with the majority of mapped reads falling in intronic regions and predominantly at the 3′ ends of transcripts (**Supplementary Fig. 1**), consistent with poly(A) capture of both mRNA transcripts and pre-mRNAs, which are abundant in nuclei[22]. In comparison with whole-cell RNA-seq methodologies (**Supplementary Fig. 4**), nuclear and whole-cell Drop-seq[17] showed shallow coverage but yielded highly similar results in terms of median transcript or unique molecular identifier (UMI) counts and gene-detection rates. Whereas nuclear and whole-cell expression levels have proven highly consistent[23], data for nuclei did show a systematic bias for longer genes (**Supplementary Fig. 4**), probably reflecting differential transcript processing and export rates associated with genic length and intron fraction[22]. Overall, we detected a median of 928 unique transcripts and 719 genes per nucleus (**Supplementary Fig. 4**), a depth expected to resolve effectively both cell-type diversity and gene expression dynamics given an increased sampling size[24]. Indeed, analysis of transcriptional heterogeneity in our data (Online Methods) resolved 35 distinct cellular clusters, including excitatory (Ex) and inhibitory (In) neuronal subtypes in the cortex[4]; distinct cerebellar granule (Gran) cells and Purkinje (Purk) neurons; and non-neuronal cells, including endothelial (End) cells, smooth muscle cells or pericytes, astrocytes (Ast), oligodendrocytes (Oli) and their precursor cells (OPCs), and microglia (Mic) (**Fig. 1b**, **Supplementary Fig. 3**). We also resolved regional differences in these populations, including cerebellar-specific Ast (Ast_Cer) and OPCs, as well as different Ex neuronal populations detected between the visual and frontal cortices (**Fig. 1c**).

To identify corresponding regulatory signatures, we generated scTHS-seq data from 32,869 single nuclei, and we resolved 27,906 of those nuclei from the visual (13,232) and frontal (4,753) cortices and the cerebellum (9,921) into neuronal or non-neuronal cell types after clustering each region independently (**Supplementary Table 2**). Of those, 15,786 data sets could be further resolved into combined cell type profiles (**Fig. 1d**). Overall, we identified 287,381 peaks associated with DNA-accessibility regions covering 144 million bp showing unique genomic alignments. This gave a median of 10,168 unique reads per cell that were confirmed to represent accessible regions (**Supplementary Fig. 5**, **Supplementary Table 1**). Analysis of human–mouse species mixing confirmed a low proportion of doublets at rates that were expected from combinatorial indexing protocols[11] (**Supplementary Fig. 5**). To identify epigenetically distinct subpopulations in the scTHS-seq data, we used an unbiased clustering strategy that modeled the probability of observing reads from a genomic site in each cell as a censored Poisson process (Online Methods). This approach accounted for the fact that the scTHS-seq signal from even the most accessible site will saturate after only a few reads.

Characterization of the identity of epigenetically defined subpopulations is more challenging than in the case of transcriptionally defined subpopulations, because functional roles of most regulatory sequences remain poorly annotated. However, on the basis of the functional annotation of the genes neighboring differentially accessible sites, we were able to distinguish broad cell types across the cortical and cerebellar regions representing Ex, In, and Gran neurons, as well as Ast, Oli, OPCs, Mic, and End cell populations (**Fig. 1d**, **Supplementary Fig. 6**). Thus, our accessibility data resolved epigenetic signatures associated with the major cell types common between frontal and visual cortices, as well as a previously undescribed neuronal signature specific to cerebellar Gran neurons (**Fig. 1e**).

Cell type and regional heterogeneity from snDrop-seq data

Using transcriptome data to define and characterize cellular identities in the different brain regions, we found expected expression of cell-type or cell-subtype marker genes (**Fig. 2a**,**b**, **Supplementary Table 3**) and profiles that were highly consistent with pooled cell populations from mouse[25] and human (temporal lobe)[26] cerebral cortex (**Supplementary Fig. 7**). Comparison with single-cell data generated from mouse visual cortex[27] and human temporal lobe[28] further confirmed broad cell-type classification and consistency between nuclear and whole-cell data (**Supplementary Fig. 7**). However, we observed over-representation of neurons at the expense of non-neuronal types such as Ast and End cells (**Supplementary Fig. 7**). Therefore, it is likely that some technical biases remain in the estimation of cell-type proportions from snDrop-seq studies. This may stem from a bias in sample processing or uneven detection rates for the cell types with lower total transcription levels (**Fig. 2a**). Ex and In cell subpopulations were annotated on the basis of correlation values with subtypes previously identified from SNS in six cortical regions[4] (**Supplementary Fig. 7**). In addition to the high correspondence, snDrop-seq permitted finer resolution of these into subpopulations (for example,

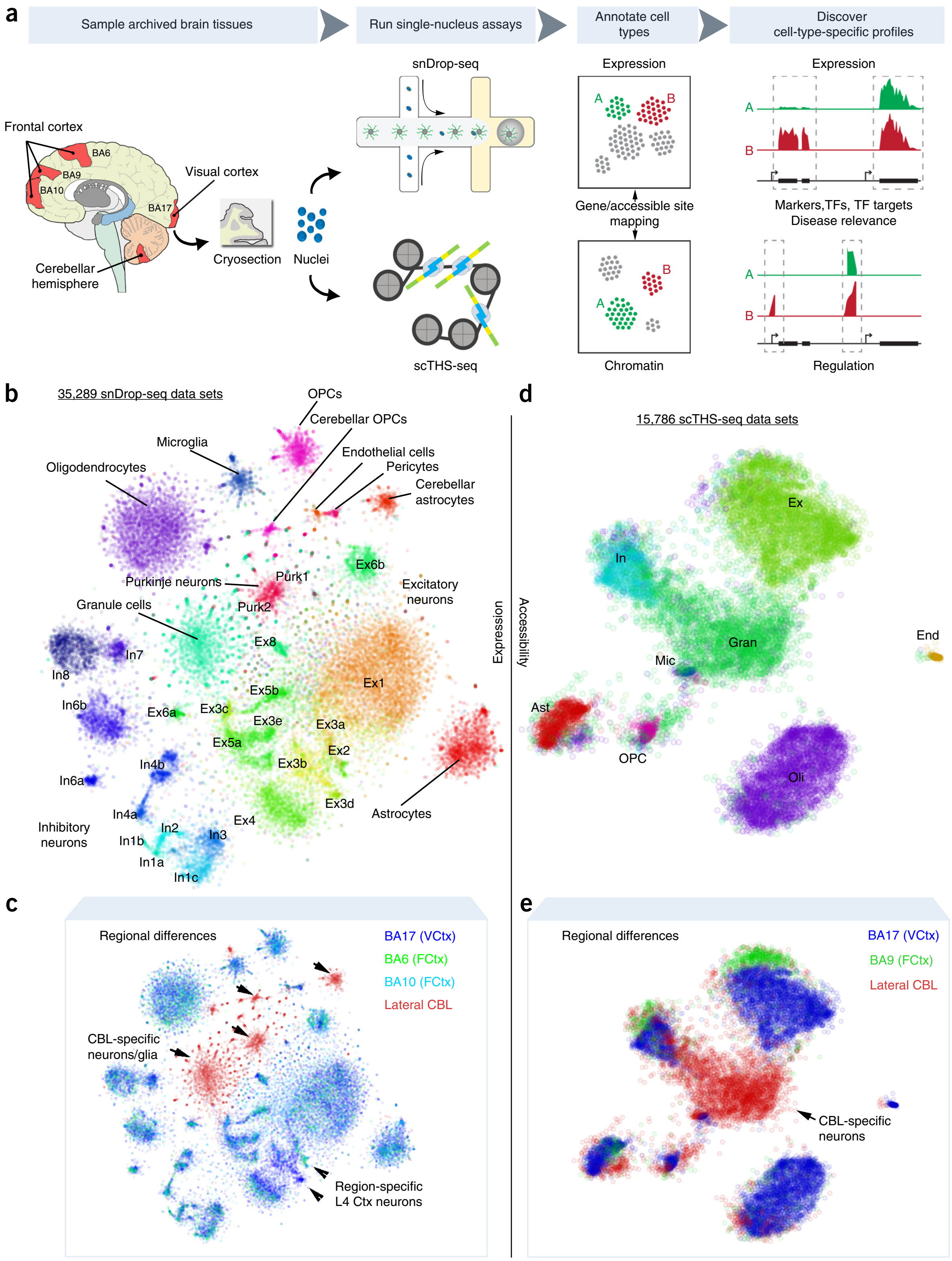

Figure 1 Integrative single-cell analyses resolve intra- and inter-regional cellular diversity in the adult human brain. (**a**) An overview of single-nucleus isolation from the visual cortex (BA17), frontal cortex (BA6, BA9, BA10), and cerebellum for snDrop-seq, scTHS-seq, and downstream expression/regulation analyses. (**b**) Combined expression (snDrop-seq) data (6 individuals, 20 experiments; **Supplementary Table 1**) showing distinct cell-type and subtype clustering visualized by *t*-distributed stochastic neighbor embedding (t-SNE). (**c**) The regional origination of data sets shown in **b**. (**d**) Combined chromatin accessibility (scTHS-seq) data (3 individuals, 3 experiments; **Supplementary Table 1**) showing the major cell-type clusters visualized (**Supplementary Table 2**) by t-SNE. (**e**) The regional origination of data sets shown in **d**. Ctx, cortex; VCtx, visual cortex; FCtx, frontal cortex; CBL, cerebellum.

Ex cell subpopulation 3 (Ex3) to Ex3a–d of the visual cortex). This demonstrates the high sensitivity of snDrop-seq in resolving neuronal subtype diversity through shallow profiling of a larger cell cohort compared with that in our previous SNS efforts[4].

We resolved Ex and Gran neurons marked by expression of *SLC17A7* and *GRM4* (**Fig. 2a**) into 14 distinct subtypes that showed enriched marker gene profiles (**Supplementary Fig. 8**, **Supplementary Table 3**) and could be distinguished by their spatial orientation[29] (**Fig. 2b**). In addition to resolving further subpopulations located in cortical layers, including the distinct *HS3ST5*$^-$*PCP4*$^+$ (Ex5a), *HS3ST5*$^+$*PCP4*$^-$ (Ex5b), and *HTR2C*$^+$*PCP4*$^+$*TLE4*$^+$ (Ex6a) subpopulations in layer 5, with the last bordering on an *HTR2C*$^-$*TLE4*$^+$ (Ex6b) layer 6 population (**Fig. 2c**), we were also able to resolve substantial regional heterogeneity among layer 4 Ex neurons, with clear expansion in the number of visual-cortex-specific subtypes (**Fig. 2a**), including the *RORB*$^+$*PCP4*$^+$ Ex3b, *RORB*$^+$*NEFM*^{hi} Ex3c, and *RORB*$^+$*PHACTR2*$^+$*EYA4*$^+$ Ex3d subpopulations (**Fig. 2c**). We further confirmed that *EYA4*$^+$ Ex3d neurons were specific to layer 4 of the visual cortex (**Fig. 2d**), but not to the frontal cortex (**Supplementary Fig. 8**). We resolved In and Purk neurons, marked by shared expression of *GAD1* (**Fig. 2a**), into 13 subtypes with enriched marker gene expression (**Supplementary Fig. 8**, **Supplementary Table 3**) that showed distinct profiles of canonical interneuron markers (e.g., *VIP*, *RELN*, *PVALB*, *SST*), as well as subtype-restricted expression (e.g., *THSD7B*, *CA8*, *GLCE*) (**Fig. 2b**). We were further able to resolve spatially distinct In neuron subpopulations, including layer 1 *RELN*$^+$*CCK*$^+$*CNR1*$^+$ In1a; upper layer *VIP*$^+$*CALB2*$^+$*TAC3*$^+$ In1d; *PVALB*$^+$*CA8*$^+$ In6a concentrated around layer 4, as well as the more peripheral *PVALB*$^+$*TAC1*$^+$ In6b; and two distinct SST-positive subtypes, the upper layer *SST*$^+$*CALB1*$^+$ (In7) and lower layer *SST*$^+$*CALB1*$^-$ (In8) subpopulations (**Supplementary Fig. 8**).

In the cerebellum, which shows a distinct cytoarchitecture compared with that of the cerebral cortex (**Fig. 2e**), we resolved multiple major cell populations, including the Gran and Purk neurons, and their supportive cell types (**Fig. 2f**,**g**). Notably, we found two distinct Purk neuron populations that expressed the inhibitory marker(s) *GAD1*/*GAD2* (**Fig. 2f**, **Supplementary Fig. 9**) and could be distinguished on the basis of *SORCS3* expression (**Fig. 2h**). Our expression data also identified two populations of Ast known to exist in this region: the velate Ast, which show transcriptomic signatures resembling those of cortical Ast and play a supportive role for Gran neurons, and Bergmann glia (Ast_Cer), which represent specialized Ast that have important roles in the laminar development of the cerebellum and which support and modulate synaptic activities of Purk neurons[30] (**Fig. 2f**). Consistently, Ast_Cer, marked by expression of ALDH1A1 (**Fig. 2g**), showed enriched expression of the AMPA-receptor-encoding gene *GRIA1* and *SLC1A3* (also known as *GLAST*), characteristic of Bergmann glia[31]. In addition to this, we resolved two distinct populations of OPCs: a *LUZP2*$^+$*CASK*$^+$ population that showed a general transcriptomic signature resembling that of the cortical OPCs, and an *ORAOV1*$^+$*LRP6*$^+$*RCN2*$^+$ population found specifically in the cerebellum (**Fig. 2f**,**g**, **Supplementary Fig. 9**). This is consistent with the majority of the cerebellar Oli originating from outside the cerebellum, and only a minority being derived from local progenitors[30]. Additional morphologically distinct cell types have been found in the cerebellar hemisphere[32] that were not resolved in this study, probably because of their more limited quantities and the oversampling of Gran neurons (**Supplementary Table 2**), which represent the most abundant cell type in this tissue. However, we did demonstrate extensive cellular expression profiling and subtype resolution by snDrop-seq for both cortical and cerebellar regions in human postmortem tissues.

An integrated transcriptome and accessibility model

To establish a more precise correspondence between the transcriptional and epigenetic states of different subpopulations, we sought to identify cells corresponding to transcriptional subpopulations in the chromatin accessibility data and cells corresponding to epigenetic subpopulations in the transcriptional data. To do so, we trained a gradient-boosting model (GBM) to predict differentially accessible genomic sites on the basis of differential expression patterns, and a separate GBM to predict differential expression on the basis of differential accessibility (**Fig. 3**), using features that included the distance of a site to a gene and the degree of differential expression or accessibility of the site or gene (Online Methods). Although the ability to predict differential expression or differential accessibility of any individual gene or site was limited, joint consideration of multiple genes or sites allowed for confident cell-type classifications (**Supplementary Fig. 10**).

Given the higher resolution of the transcriptional data, we sought to use this model to further partition chromatin-accessibility clusters by identifying chromatin-accessibility signatures associated with the observed transcriptional subpopulations (**Fig. 3a**). Briefly, using transcriptional data, we first carried out hierarchical clustering of the identified cell types on the basis of their cumulative expression signature to establish a cell-type relationship dendrogram. We then performed iterative binary splits on this dendrogram and identified genes that were differentially expressed between the two branches. We then applied our GBM classifier to predict differentially accessible genomic sites. We used joint consideration of all predicted differentially accessible sites to classify the cells measured by scTHS-seq as corresponding to either branch on the basis of the pattern of their chromatin accessibility. Using the results from this initial classification, we built a refined differential chromatin accessibility signature and used it to determine the final branch assignment and assess the stability of the branch annotations through cross-validation (Online Methods).

In this way, we used transcriptional data to identify genes that were differentially expressed between non-neuronal and neuronal cell types. The predicted differentially accessible sites allowed us to confidently resolve non-neuronal and neuronal cell types in the chromatin-accessibility data. Having resolved neuronal cell types, we repeated the procedure to distinguish Ex from In cells. Then we used the procedure to resolve different In neuron subtypes (**Fig. 3b**,**c**), and so on. In this manner, we were able to identify epigenetic differences relevant to In neuron subtypes (InA, InB) distinguished by their developmental origin from subcortical regions of the medial or lateral/caudal ganglionic eminences[4,33,34] (**Fig. 3b**). However, attempts to further resolve additional In subtypes in InA and InB populations resulted in low stability of cell identities, suggesting a lack of differentially accessible sites sufficient to consistently distinguish between the two predicted groups (**Fig. 3b**,**c**). Similarly, although layer 4 Ex neurons (ExL4, comprising Ex2–4) could not be distinguished from ExL5 and ExL6 neurons (Ex5–8) on the basis of an unbiased analysis of chromatin accessibility data alone, the integration of differential expression information from the higher-resolution transcriptional data allowed us to identify relevant differentially accessible sites to further partition chromatin-accessibility clusters (**Fig. 3d**–**f**). We were able to confidently resolve all major cell types as expected, such as Oli, OPCs, Ast, End cells, In cells, and Ex cells from the visual cortex (**Fig. 3c**); Ast, Oli, In cells, and Ex cells in the frontal cortex; and Gran cells, Oli, and End cells in the cerebellum (**Supplementary Fig. 6**). We further confirmed that the resolved cell types and subtypes showed enriched accessibility at promoters of marker genes (**Fig. 3g**,

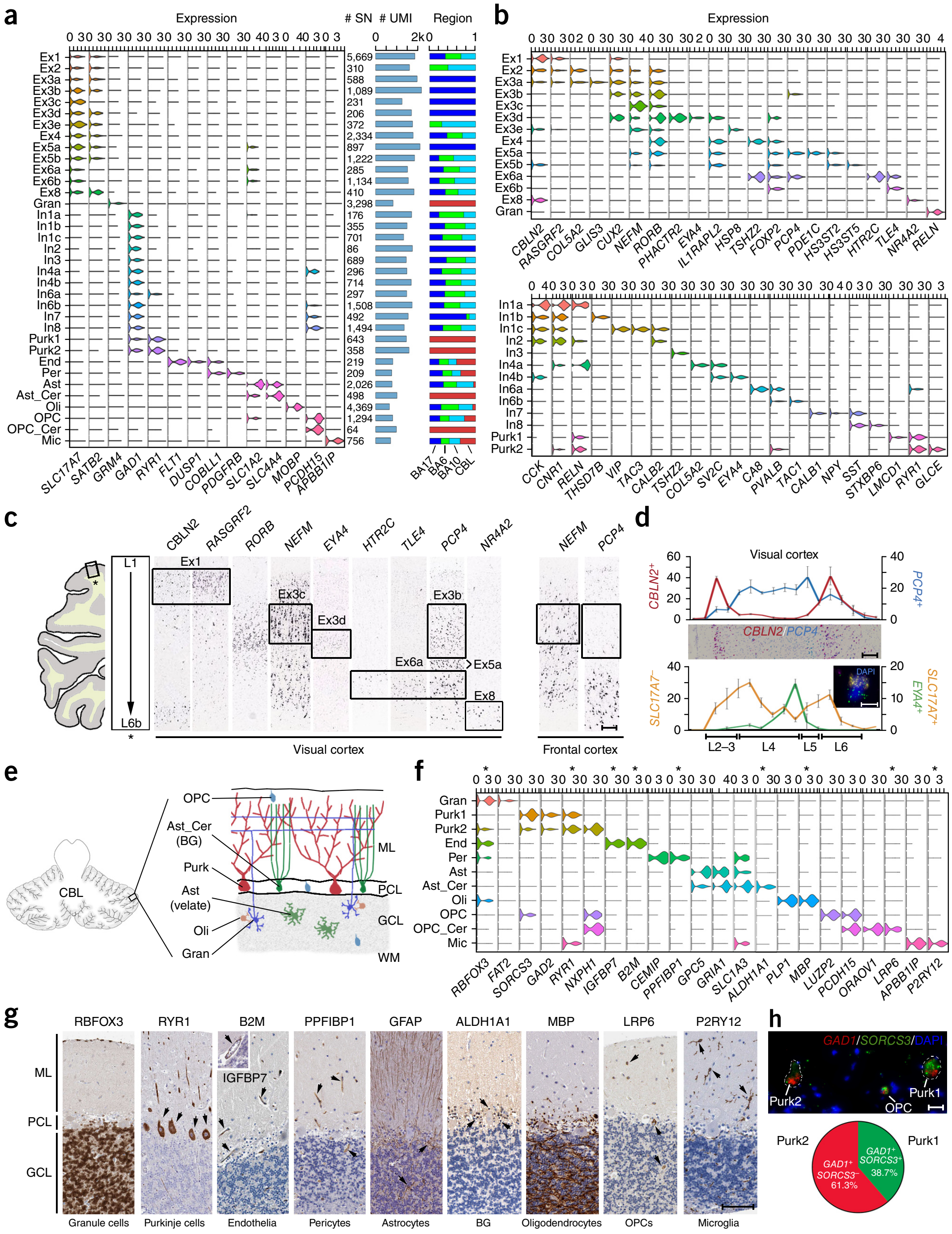

Figure 2 Expression data permit the identification and classification of molecularly and spatially distinct cell types and subtypes. (**a**) Violin plots of expression values for cell-type-specific marker genes. The number of data sets (SN), average transcript (UMI) count, and relative proportion across regions sampled (**Fig. 1c**) are indicated for each cluster. (**b**) Top, gene expression values of layer-specific[4,29] and subtype-enriched markers for excitatory neuronal subtypes. Bottom, expression values for classical interneuron marker genes[4] and subtype-enriched transcripts. (**c**) RNA *in situ* hybridization (ISH) stains (Allen Human Brain Atlas[58]; **Supplementary Table 9**) of the visual cortex for select marker genes from **b**. Frontal cortex stains demonstrated the absence of associated layer 4 subpopulations. Scale bar, 200 μm. (**d**) Top, RNA ISH counts showing the number of cells positive for *CBLN2* and *PCP4* (chromogenic image shown; scale bar, 200 μm) in image fields spanning from the pial layer to the white matter. Bottom, RNA ISH counts for *SLC17A7*-positive cells and for cells positive for both *SLC17A7* and *EYA4* (inset; scale bar, 10 μm). The data shown are mean values for four separate layer cross-sections (replicate regions) from a single individual, ±s.d. (**e**) A schematic of the cerebellar (CBL) cytoarchitecture. ML, molecular layer; PCL, Purkinje cell layer; GCL, granule cell layer; WM, white matter. BG, Bergmann glia. (**f**) Violin plots of expression values for type-specific marker genes specifically for cerebellar data. Asterisks indicate markers shown in **g**. (**g**) Protein staining (Human Protein Atlas[59]; **Supplementary Table 10**) for select cell-type-specific markers indicated by asterisks in **f**. Scale bar, 100 μm. (**h**) Top, fluorescent RNA ISH image (adjusted for visualization; Online Methods) showing representative *GAD1*$^{+}$ Purk1 (*SORCS3*$^{+}$) and Purk2 (*SORCS3*$^{-}$ or *SORCS3*^{lo}) neurons. OPCs showing low expression of *GAD1* were also *SORCS3*$^{+}$. Scale bar, 20 μm. Bottom, the proportions of *GAD1*$^{+}$/*SORCS3*$^{+}$ and *GAD1*$^{+}$/*SORCS3*$^{-}$ populations quantified from imaged Purk neurons from a single individual (**Supplementary Fig. 9d**). Per, pericytes; OPC_Cer, cerebellum-specific OPCs.

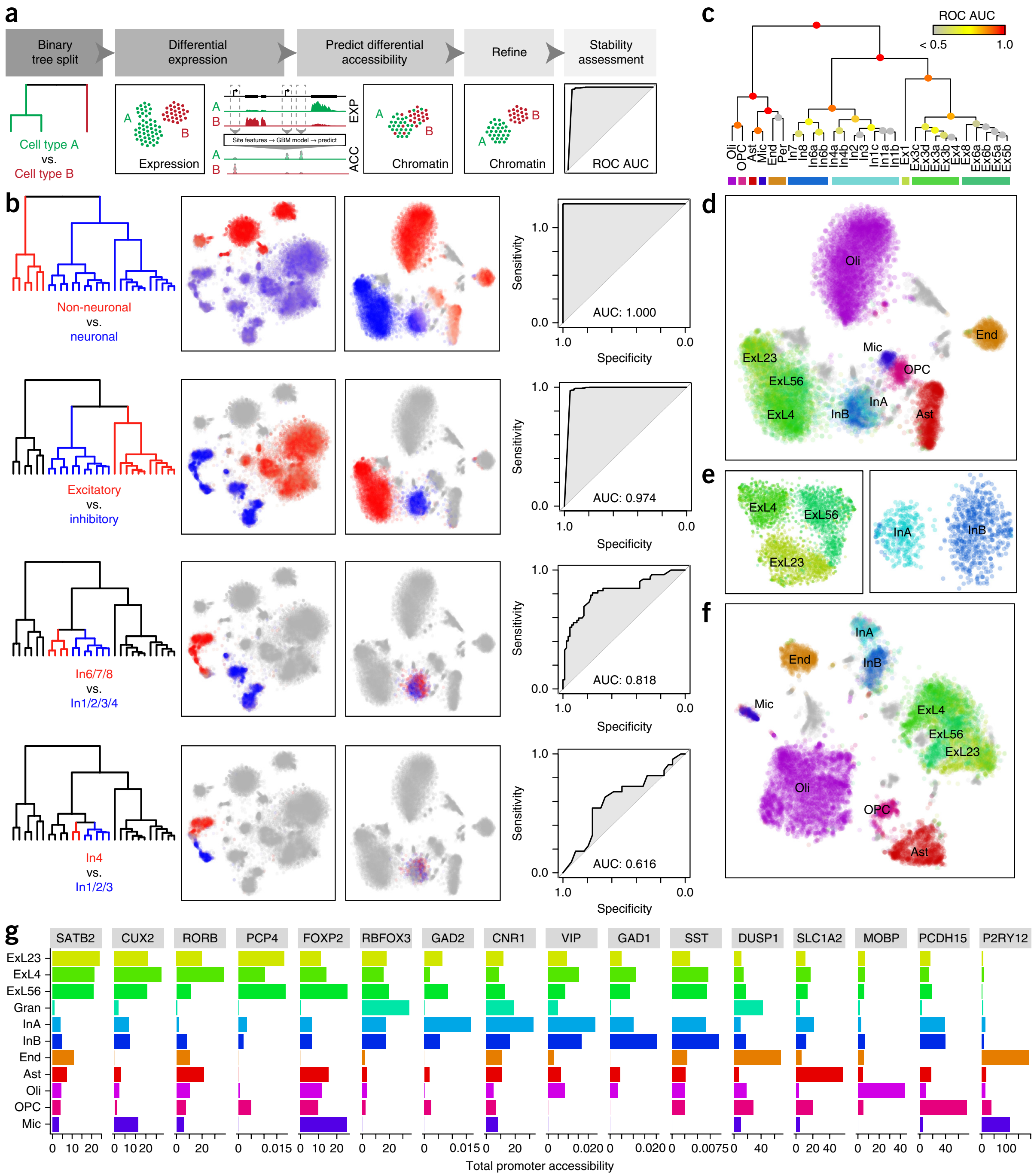

Figure 3 Integrative mapping of transcriptional and epigenetic subtypes. (**a**) Overview. First, a taxonomy of cell types is constructed on the basis of the expression data. For each binary split in the transcriptional taxonomy, a set of genes that are differentially expressed between the two branches is identified. A GBM is used to predict a set of differentially accessible chromatin sites corresponding to the identified differential expression signature, to classify scTHS-seq cells as belonging to either branch. Predicted branch annotations are refined via the identification of differentially accessible sites from scTHS-seq data. The stability of the branch annotations is assessed by cross-validation (Online Methods). ROC, receiver operating characteristic; AUC, area under the curve. ACC, accessibility; EXP, expression. (**b**) Identification of inhibitory neuron subpopulations via the integrative approach, using visual cortex data (snDrop-seq, 5 individuals; scTHS-seq, 1 individual). In the top binary split of transcriptional taxonomy, neuronal cells are separated from non-neuronal cells. Differentially expressed genes ($Z > 1.96$) are identified. The average expression of significantly upregulated genes ($P < 0.05$, Wilcoxon test) in each branch is shown, with red corresponding to high expression in the red branch and blue corresponding to high expression in the blue branch. Predicted differentially accessible sites are visualized in the same way. Prediction performance, as assessed on the basis of ROC curves and AUC, demonstrated high stability of split for non-neuronal versus neuronal, Ex versus In, and In1/2/3/4 versus In6/7/8, but not In4 versus In1/2/3. (**c**) A summary of stability for each binary split of the transcriptional taxonomy. (**d**) Final cell type predictions from the integrated analysis projected onto the original visual cortex scTHS-seq data t-SNE embedding. (**e**) Refinement of the visual cortex scTHS-seq data t-SNE embedding for Ex (left) and In (right) subpopulations only, integrating predicted differentially accessible sites. (**f**) Refinement of the complete visual cortex scTHS-seq data t-SNE integrating predicted differentially accessible sites. (**g**) The accessibility of select marker genes. Read-mapping to promoters of each gene for all cells in each epigenetic subpopulation from **f** was averaged for the number of sites and cells for comparison across subpopulations.

Supplementary Table 4). Thus, despite the lower intrinsic cell-type resolution of accessibility data compared with that of transcriptional data, computational integration of both scTHS-seq and snDrop-seq results allowed us to reconstruct detailed epigenetic profiles of fine-grained cell types in the brain, enabling investigations of the regulatory processes active in each cell type.

Transcription factor activities in remyelination

Having established the cell-type identity of each epigenetically distinct subpopulation, we sought to identify transcription factors (TFs) relevant to the regulatory states of each cell type. To do so, we looked for TFs whose predicted binding sites were over-represented in the regions of differential chromatin accessibility that distinguished a given cell type (**Fig. 4a**). By screening a set of 379 TFs with position weight matrices from the JASPAR database[35], we identified TFs that showed statistically significant association with at least one of the major cell types in the visual cortex (**Fig. 4b**, **Supplementary Table 5**). We further found TF activities potentially specific to spatially distinct Ex neuron subpopulations (L2/3 versus L4 versus L5/6), as well as to In neurons derived from different subcortical regions (**Supplementary Table 5**). As an independent validation, we cross-validated with snDrop-seq data to confirm that the TFs that showed significant association with a particular cell type or subtype also tended to have higher expression levels in that cell type (**Supplementary Fig. 11**).

To further demonstrate the utility of an integrative approach for uncovering relevant biology, we focused on the transition of OPCs to Oli in the adult brain. Myelin regeneration occurs through neuronal activation and differentiation of OPCs into myelinating Oli that resheath neuronal axons to restore saltatory conduction and normal functionality. Dysregulation of this process can lead to severe neurologic disorders, including multiple sclerosis[36,37]. Notably, we found specific TF signatures that distinguished OPC and Oli populations (**Fig. 4b**, **Supplementary Table 5**). To determine whether these could reveal key regulatory processes underlying adult remyelination, we assessed differentiation states and associated gene expression signatures in these lineages in the visual cortex. Using diffusion mapping with Destiny[38], we oriented OPCs and Oli along a developmental trajectory (**Fig. 4c**, **Supplementary Fig. 12**) and assessed differential expression among cells in the beginning, middle, and end of that trajectory. In doing so, we found intermediary cells (immature Oli (iOli)) with a unique expression signature independent of experimental batch (**Fig. 4c,d**, **Supplementary Fig. 12**, **Supplementary Table 6**) that could provide insight into the early mechanisms of human adult Oli maturation. In agreement with findings in the mouse[39], our human OPC population expressed marker genes associated with mouse OPCs (*PDGFRA*, *CSPG4*, *SOX6*, *VCAN*), yet it also expressed markers for more committed mouse progenitors (*ITPR2*, *NEU4*), indicating the inability to distinguish these subtle states in our human data (**Fig. 4d**, **Supplementary Table 6**). Furthermore, the mature Oli (mOli) population expressed markers associated with myelin formation (*PLP1*, *MBP*, *MOG*) (**Fig. 4d**, **Supplementary Table 6**) and did not resolve into further subpopulations as seen in the mouse, which might account for the absence of juvenile states in the adult human brain.

However, the progressive expression signatures found in OPCs, iOli, and mOli, which were conserved across brain regions and independent of the ordering method, could be further refined into stages of an OPC glutamate-activation response[40] (**Supplementary Fig. 13**). Recent studies have proposed that AMPA and kainate receptors mediate an initial axon–OPC synaptic response to glutamate that directs OPCs to exposed axonal sites where NMDA-receptor activation directs remyelination[40–44]. In agreement with this finding, our data showed

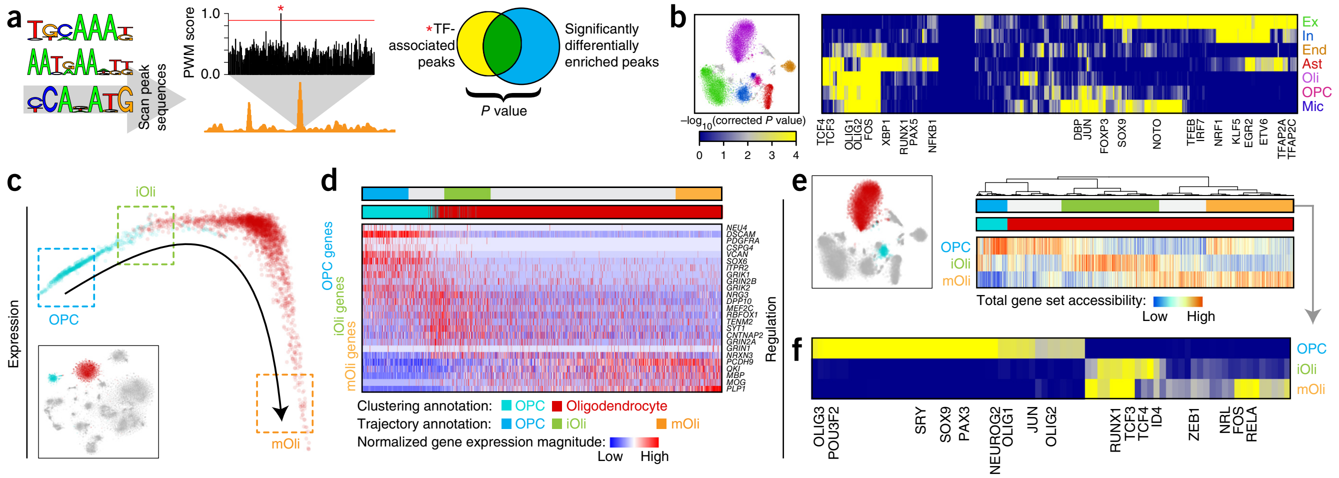

Figure 4 Mapping of transcription factor (TF) activities to specific cell types to resolve remyelination programs. (**a**) A schematic of TF analysis. Briefly, putative TF binding sites (TFBSs) were identified within all hypersensitive sites on the basis of matching position weight matrices (PWMs). To identify relevant factors for a given cell type, we tested sites that showed significant differential accessibility ($P < 0.05$) in that cell type for enrichment of different TFBSs by Fisher's exact test (corrected $P < 0.2$; see Online Methods). (**b**) A heat map of TF association with epigenetic subpopulations (right). Each column represents a TF, and each row represents an epigenetic subpopulation (left) from the visual cortex scTHS-seq data (1 individual). Select TFs are annotated. (**c**) A diffusion map pseudotime trajectory for 3,064 snDrop-seq data sets (644 OPC and 2,420 Oli) from the visual cortex (5 individuals) obtained with Destiny (inset). Data are color-coded according to the original data set annotations from clustering analysis. Refined annotations based on the inferred pseudotime trajectory are indicated by boxes. (**d**) A heat map of selected genes involved in the remyelination program. Columns represent data sets ordered by the pseudotime trajectory in **c**. Rows represent genes ordered by association with OPCs, iOli, and mOli on the basis of the significance of differential upregulation in each group. (**e**) Left, the accessibility of genes involved in remyelination programs for OPC and Oli scTHS-seq data sets from the visual cortex. Right, a heat map of total promoter accessibility. Each column represents a cell, and rows represent accessibility for genes differentially upregulated in OPCs, iOli, and mOli, respectively. (**f**) A heat map of TF association with stages of Oli maturation. Each column represents a TF, and each row represents an epigenetic subpopulation inferred from **e**. Select TFs are annotated.

that genes encoding AMPA receptor and kainate receptor (*GRIN* and *GRIK4*, respectively) were enriched in OPCs and iOli, and that NMDA-receptor-encoding genes (*GRIN2A*, *GRIN2B*) were enriched only in the iOli subpopulation (**Fig. 4d**, **Supplementary Table 6**). Functional ontogenies for gene sets identified in OPC maturation progressed along six stages: (i) neurogenesis (progenitor marker expression), (ii) glutamate-receptor activities, (iii) synaptic transmission, (iv) ion channel activities, (v) membrane assembly, and (vi) axon ensheathment (**Supplementary Fig. 13**). These results provide independent support for mechanisms of neuronal activity in remyelination.

To understand the regulatory mechanisms that define these gene expression dynamics, we assessed the accessibility of the differentially upregulated genes in OPC and Oli subpopulations in the visual cortex from our scTHS-seq data. In agreement with our expression data, regulatory sites for OPC, iOli, and mOli gene sets showed differential accessibility (**Fig. 4e**). Further, OPC and iOli gene accessibilities showed nearly complete mutual exclusivity, indicating active regulatory mechanisms that might maintain these two states. Most significant TF activities within OPC accessible sites were associated with SOX9 (**Fig. 4f**, **Supplementary Table 7**), which is known to be required for mouse OPC specification[45], survival, and migration[46]. Moreover, we found that the iOli-specific accessible sites showed significant enrichment for TCF4 TF binding (**Fig. 4g**, **Supplementary Table 7**), which has an important role in modulating Wnt/β-catenin to promote remyelination in the mouse[47,48]. Thus, our TF analyses implicate conserved regulatory mechanisms that maintain adult oligodendrocyte progenitors and coordinate their maturation for remyelination.

Mapping of pathogenic risk to specific brain cell types

Cell-type-specific epigenomic information has been highly valuable for identifying pathogenic cell types and specific regulatory mechanisms underlying many common genetic diseases[49–51], yet brain diseases remain inadequately understood. To address this, we used data from the US National Institutes of Health GRASP catalog (https://grasp.nhlbi.nih.gov/Search.aspx) for significant single-nucleotide polymorphisms (SNPs; $P < 10^{-6}$) identified from genome-wide association studies (GWASs) for ten brain-related disorders, as well as SNPs for seven non-brain-related diseases as controls. Given that causal variants are often located at different positions in linkage disequilibrium with GWAS SNPs, we searched for enrichment of DNA-accessibility regions in 100-kb windows centered on all GWAS SNPs of a given disease, and assessed the significance by random permutations (**Fig. 5a**, Online Methods). This analysis identified strong disease-specific enrichments in multiple cell types and subtypes, contrasting with an alternative possibility of uniformity (**Fig. 5b–d**, **Supplementary Fig. 14**, **Supplementary Table 8**). Notably, we found highly significant enrichment for Alzheimer's disease (AD) risk variants in Mic (Z-score = 5.41; **Supplementary Table 8**), which is in line with the significant Mic signatures found to be activated in late-onset AD cortex[52] and for AD risk variants that show higher expression in Mic[53]. Comparisons with bulk ATAC-seq data[53] demonstrated the sensitivity of our single-cell data to predict Mic regulatory sites and their associated disease-specific risk variant enrichments (**Fig. 5e–g**, **Supplementary Table 8**). We did not find any significant enrichments for non-brain-related disease variants in neurons, which further supports our predictions for disease-pathogenic cell types. In fact, a majority of non-brain-related enrichments were for cell types closely related to those implicated in these diseases, such as Mic and End cells in autoimmune diseases (Crohn's, celiac, and type 1 diabetes). Thus, our single-cell regulatory maps were highly consistent with bulk studies and may permit linkage to cell-type-specific disease risks. Although further validation involving much larger samples sizes, other disease data sets, and mechanistic studies should be pursued in the future, our chromatin maps provide a cell-type- or cell-subtype-specific data set through which new aspects of brain diseases can be understood.

DISCUSSION

Reconstruction of cellular composition is important for improving understanding of both normal functions of the human brain and mechanisms of dysfunction and disease. This study demonstrates a large-scale, integrative transcriptomic and epigenomic single-cell analysis of the human adult brain, using two highly scalable methods applicable to postmortem tissues: snDrop-Seq and scTHS-seq. Using nuclei isolation to overcome challenges associated with live tissues or the processing of archived tissues, we recovered 35 subpopulations of non-neuronal and neuronal cell types in human adult cortex and cerebellar hemisphere. Our results underscore the power of sparse sampling of single cells in complex tissues at a massive scale: as long as the data from individual cells are informative enough for clustering and 'virtual sorting' into different groups[54,55], they can be combined into aggregate profiles that are as rich as bulk sequencing of different cell populations. This also applies to accessibility data, accounting for the greater cell-type discovery observed for the larger visual cortex data set (**Supplementary Fig. 6**). However, despite the increased coverage allowed by scTHS-seq, chromatin-accessibility data, on its own, showed less power to resolve finer cellular subtypes, reflecting the need for improvements in sensitivity. Further, although snDrop-seq permits more wide-ranging tissue profiling than our previously published method[4], we were unable to distinguish subpopulations of cortical Ast and Oli found in mouse studies[39,56]. It remains unclear whether this might be attributed to technical artifacts associated with nuclear isolation, the more limited detection of transcripts in glia, differences in the tissues or regions sampled, differences associated with tissue archiving, or biologically limited heterogeneity in the mature adult human brain. However, our expression data extensively resolved neuronal and non-neuronal subpopulations, as well as distinct subtypes between the cerebral and cerebellar cortices. Furthermore, using our combined transcriptomic and epigenomic profiles, we were able to detect evolutionarily conserved expression and regulation dynamics underlying adult remyelination, thus demonstrating the sensitivity of our methods to resolve the cellular heterogeneity and genetic programs that exist in the adult human brain.

We have additionally outlined a computational strategy for mapping between corresponding transcriptional and epigenetic states that can be used to reconstruct aggregate epigenetic profiles for fine-grained cell types. Such profiles provide valuable insights into the regulatory processes and elements that shape the identity of different cell types, as well as their relevance to human disease. Whereas previous studies identified pathogenic cell types for several common human diseases, our analysis provides proof-of-concept data that can be used to assess common genetic risk alleles in multiple cell types of an organ, particularly the brain. It provides a coherent framework to consolidate previous GWAS findings, such as the relative contributions of glia, Mic, and neurons to sporadic AD[57], and could potentially extend to single-neuron genomic mosaicism that also becomes altered in this disease[14]. Generating multiple types of -omics maps from single cells en masse leverages the strength of each method to synergistically increase the confidence of cell-type assignment to enrich cell annotations. This combined approach thus represents a strategy for the systematic construction of atlases composed of single-cell data for human organs such as the brain and, eventually, for the full human body.

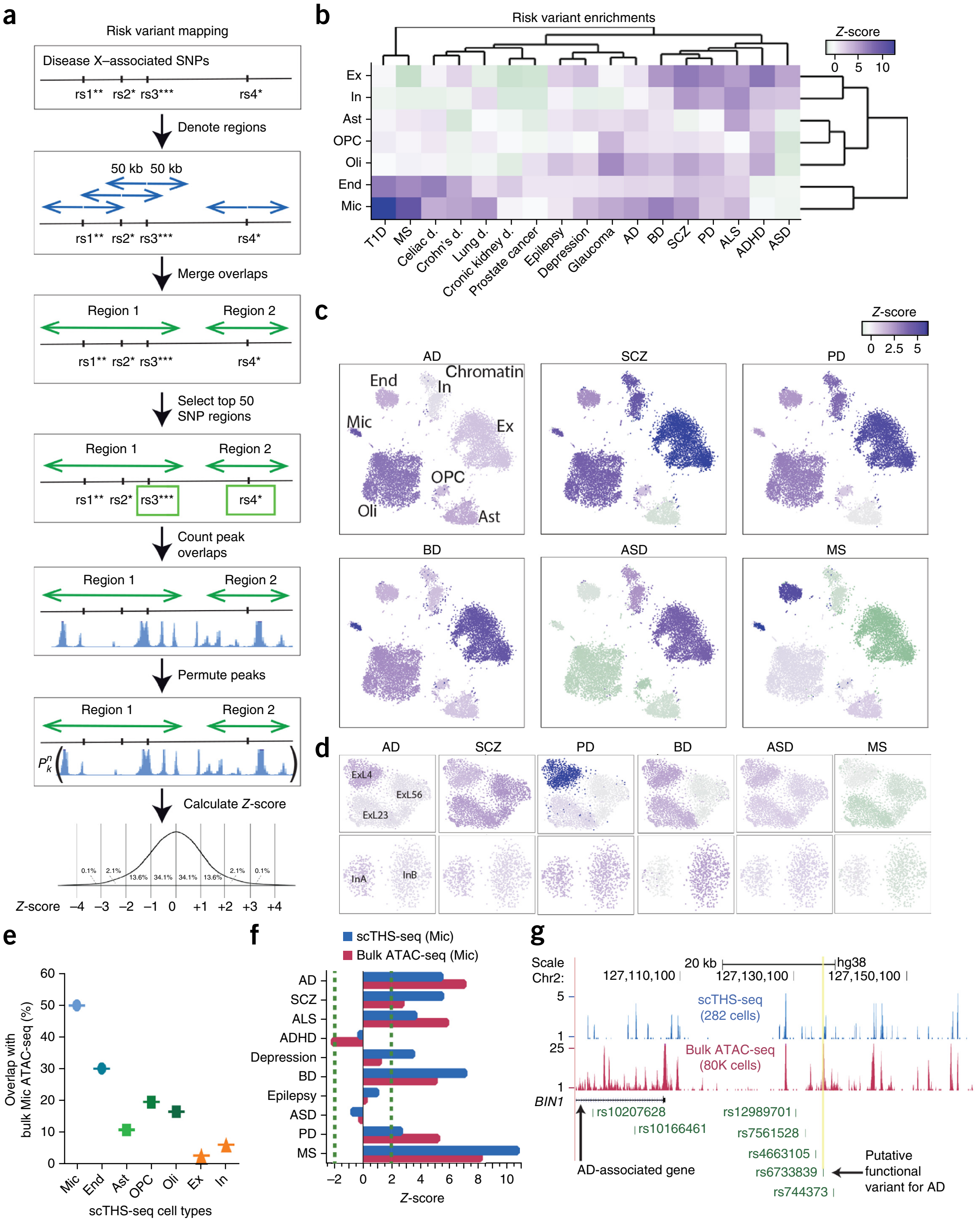

Figure 5 Mapping of common disease risk variants to specific brain cell types. (**a**) Method overview. Briefly, GWAS SNPs were obtained for each disease, extended to 100 kb, and merged. Then the top 50 most significant SNPs were selected, the number of peaks in overlaps was counted, peaks were permuted, the number of peaks was counted in each region for each permutation, and finally *Z*-scores were calculated. Asterisks indicate the relative significance level. (**b**) A heat map representing the enrichment *Z*-scores across seven cell clusters (rows) identified from visual cortex scTHS-seq data (1 individual) for ten brain diseases (columns) and seven unrelated diseases (**Supplementary Table 8**). T1D, type 1 diabetes; MS, multiple sclerosis; d., disease; AD, Alzheimer's disease; BD, bipolar disorder; SCZ, schizophrenia; PD, Parkinson's disease; ALS, amyotrophic lateral sclerosis; ADHD, attention deficit hyperactivity disorder; ASD, autism spectrum disorder. Dark purple and purple indicate a significant *Z*-score (>1.96), whereas light purple, gray, and light green indicate a non-significant *Z*-score, and green represents a significant negative association (*Z*-score < −1.96). (**c**) *Z*-scores for the enrichment of GWAS SNPs in the open chromatin of Ex cell, In cell, Oli, OPC, Ast, End cell, and Mic populations were overlaid on the cell clusters. Data for six brain disorders are shown. (**d**) *Z*-scores for the enrichment of GWAS SNPs in open chromatin of three excitatory subclusters and two inhibitory subclusters. In **c** and **d**, *Z*-score color-coding is as in **b**. (**e**) The percent overlap of published bulk microglia ATAC-seq[53] data with differential peaks for each cell population identified from scTHS-seq data. (**f**) A comparison of GWAS SNP enrichment in open chromatin from published bulk microglia ATAC-seq data and differential open chromatin regions from scTHS-seq microglia data. (**g**) Visualization of combined scTHS-seq data and published bulk ATAC-seq data on microglia over the gene and promoter region of the AD-associated gene *BIN1*. The putative causal SNP for AD located in a PU.1 binding footprint[53] is also indicated.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Flow cytometry was done at both the UCSD Human Embryonic Stem Cell Core and the TSRI Flow Cytometry Core. We thank T.F. Osothprarop and M.M. He (Illumina, San Diego, California, USA) for providing Tn5059 transposase; N. Salathia for assistance in sequencing; Y. Wu, T. Pakozdi, and Z. Chiang for assistance in sequencing analysis; and G. Kennedy for help on RNAscope. Funding support was provided by the NIH Common Fund Single Cell Analysis Program (1U01MH098977 to K.Z. and J.C.). T.E.D. is a fellow in the Pediatric Scientist Development Program and is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (award no. K12-HD000850). G.E.K. was supported by a Neuroplasticity of Aging Training Grant (5T32AG000216-24). P.V.K. was supported by the NIH Centers for Excellence in Genomic Science (P50MH106933), NIH grant 1R01HL131768, and an NSF CAREER award (NSF-14-532). J.F. was supported by NIH grant F31 CA206236. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx samples used for the analyses described in this study were obtained from the University of Miami Brain Endowment Bank.

AUTHOR CONTRIBUTIONS

K.Z., P.V.K., and J.C. oversaw the study. B.B.L., S.C., B.C.S., G.E.K., Y.C.Y., T.E.D., and D.G. performed experiments. J.F., B.B.L., S.C., and B.C.S. performed bioinformatics analyses. B.B.L., S.C., B.C.S., J.F., J.C., P.V.K., and K.Z. wrote the manuscript, with input from all other co-authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Grindberg, R.V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. USA* **110**, 19802–19807 (2013).
2. Habib, N. *et al.* Div-Seq: single-nucleus RNA-seq reveals dynamics of rare adult newborn neurons. *Science* **353**, 925–928 (2016).
3. Krishnaswami, S.R. *et al.* Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons. *Nat. Protoc.* **11**, 499–524 (2016).
4. Lake, B.B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586–1590 (2016).
5. Lacar, B. *et al.* Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat. Commun.* **7**, 11022 (2016).
6. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
7. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
8. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
9. Sos, B.C. *et al.* Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* **17**, 20 (2016).
10. Buenrostro, J.D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
11. Cusanovich, D.A. *et al.* Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
12. Jin, W. *et al.* Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
13. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
14. Bushman, D.M. *et al.* Genomic mosaicism with increased amyloid precursor protein (*APP*) gene copy number in single neurons from sporadic Alzheimer's disease brains. *eLife* **4**, e05116 (2015).
15. Gole, J. *et al.* Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013).
16. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
17. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
18. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
19. Kia, A. *et al.* Improved genome sequencing using an engineered transposase. *BMC Biotechnol.* **17**, 6 (2017).
20. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
21. Corces, M.R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
22. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
23. Lake, B.B. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* **7**, 6031 (2017).
24. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
25. Zhang, Y. *et al.* An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J. Neurosci.* **34**, 11929–11947 (2014).
26. Zhang, Y. *et al.* Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron* **89**, 37–53 (2016).
27. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
28. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. USA* **112**, 7285–7290 (2015).
29. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483–496 (2012).
30. Buffo, A. & Rossi, F. Origin, lineage and function of cerebellar glia. *Prog. Neurobiol.* **109**, 42–63 (2013).
31. Saab, A.S. *et al.* Bergmann glial AMPA receptors are required for fine motor coordination. *Science* **337**, 749–753 (2012).
32. Butts, T., Green, M.J. & Wingate, R.J. Development of the cerebellum: simple steps to make a 'little brain'. *Development* **141**, 4031–4041 (2014).
33. Hansen, D.V. *et al.* Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nat. Neurosci.* **16**, 1576–1587 (2013).
34. Ma, T. *et al.* Subcortical origins of human and monkey neocortical interneurons. *Nat. Neurosci.* **16**, 1588–1597 (2013).
35. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
36. Choi, J.W. *et al.* FTY720 (fingolimod) efficacy in an animal model of multiple sclerosis requires astrocyte sphingosine 1-phosphate receptor 1 (S1P1) modulation. *Proc. Natl. Acad. Sci. USA* **108**, 751–756 (2011).
37. Groves, A., Kihara, Y. & Chun, J. Fingolimod: direct CNS effects of sphingosine 1-phosphate (S1P) receptor modulation and implications in multiple sclerosis therapy. *J. Neurol. Sci.* **328**, 9–18 (2013).
38. Angerer, P. *et al.* destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
39. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science* **352**, 1326–1329 (2016).
40. Gautier, H.O. *et al.* Neuronal activity regulates remyelination via glutamate signalling to oligodendrocyte progenitors. *Nat. Commun.* **6**, 8518 (2015).
41. Hines, J.H., Ravanelli, A.M., Schwindt, R., Scott, E.K. & Appel, B. Neuronal activity biases axon selection for myelination *in vivo*. *Nat. Neurosci.* **18**, 683–689 (2015).
42. Lundgaard, I. *et al.* Neuregulin and BDNF induce a switch to NMDA receptor-dependent myelination by oligodendrocytes. *PLoS Biol.* **11**, e1001743 (2013).
43. Mensch, S. *et al.* Synaptic vesicle release regulates myelin sheath number of individual oligodendrocytes *in vivo*. *Nat. Neurosci.* **18**, 628–630 (2015).
44. Wake, H., Lee, P.R. & Fields, R.D. Control of local protein synthesis and initial events in myelination by action potentials. *Science* **333**, 1647–1651 (2011).
45. Pozniak, C.D. *et al.* Sox10 directs neural stem cells toward the oligodendrocyte lineage by decreasing Suppressor of Fused expression. *Proc. Natl. Acad. Sci. USA* **107**, 21795–21800 (2010).
46. Finzsch, M., Stolt, C.C., Lommes, P. & Wegner, M. Sox9 and Sox10 influence survival and migration of oligodendrocyte precursors in the spinal cord by regulating PDGF receptor alpha expression. *Development* **135**, 637–646 (2008).
47. Zhao, C. *et al.* Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation. *Nat. Commun.* **7**, 10883 (2016).
48. Rocha, H., Sampaio, M., Rocha, R., Fernandes, S. & Leão, M. MEF2C haploinsufficiency syndrome: report of a new MEF2C mutation and review. *Eur. J. Med. Genet.* **59**, 478–482 (2016).
49. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
50. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
51. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

52. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
53. Gosselin, D. *et al.* An environment-dependent transcriptional network specifies human microglia identity. *Science* **356**, eaal3222 (2017).
54. Fan, H.C., Fu, G.K. & Fodor, S.P. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
55. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
56. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
57. Mattson, M.P. Pathways towards and away from Alzheimer's disease. *Nature* **430**, 631–639 (2004).
58. Hawrylycz, M.J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
59. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

ONLINE METHODS

Sample origin and nuclei preparation. All human tissue protocols were approved by the Office for Human Research Protection at Sanford Burnham Prebys Medical Discovery Institute and conformed to National Institutes of Health guidelines. Nuclei were prepared with nuclear extraction buffer (NEB) as described previously[4]. Briefly, fresh-frozen postmortem brain tissue was sectioned at 50 µm with a cryostat and placed in 1 ml of ice-cold NEB for 10 min. Nuclei were extracted with 10–12 up-and-down strokes of a glass Dounce homogenizer with a Teflon pestle in 1 ml of NEB. Samples were passed through a 50-µm filter (Sysmex Partec) and then incubated on ice for 10 min. Samples were spun for 5 min at 250–300*g*, washed in PBS + 2 mM EGTA, and resuspended in PBS + 2 mM EGTA supplemented with 1% fatty-acid-free BSA (Gemini) containing 4′,6-diamidino-2-phenylindole (DAPI). DAPI$^+$ single nuclei were purified by flow cytometry with a MoFlo Astrios (Beckman Coulter) or FACSAria Fusion (Becton Dickinson), concentrated at 900*g* for 10 min, and then used directly for droplet encapsulation.

Nuclei encapsulation, mRNA-seq library preparation, and sequencing. Drop-seq was performed as described previously[17], but with modifications optimized for nuclei processing, in a procedure now termed snDrop-seq. Before droplet generation, connecting tubing and syringes were coated with 1% BSA to prevent nonspecific binding of nuclei to the surface, and then rinsed with PBS. To reduce nuclei settling, Ficoll PM-400 was added to the nuclei suspension buffer, rather than the lysis buffer. Nuclei were loaded at a concentration of 100 nuclei/µl and coencapsulated in droplets with barcoded beads purchased from ChemGenes Corporation (cat. no. Macosko201110). When encapsulation was complete, the contents of the droplet-collecting Falcon tubes were overlaid with a layer of mineral oil and then transferred to a 72 °C water bath. After 5 min of incubation, the tubes were moved from the water bath to ice, and droplets were broken by perfluorooctanol, after which beads were harvested and hybridized RNA was reverse-transcribed. cDNA was then PCR-amplified for 16 cycles with primer, buffer, and cycle conditions identical to those described previously[17]. A total of 46 libraries were prepared from 20 experiments (**Supplementary Table 1**), and cDNA from each replicate was prepared and tagmented by Nextera XT and indexed with different Nextera index 1 primers. cDNA libraries were pooled and sequenced on an Illumina HiSeq 2500 with Read1CustSeqB[17] for priming of read 1 (read 1 was 30 bp; read 2 (paired end) was 120 bp).

snDrop-seq data processing. Paired-end sequencing reads were processed largely as described (http://mccarrolllab.com/wp-content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf), with additional correction steps. First, paired-end reads were filtered out if read 1 had more than four non-T bases in the last ten bases (to remove all non-poly(T)-captured contaminated reads), or had one or more bases with a poor quality score (<10). Cell barcode and UMI information were then inferred from the first 12 bases and the next 8 bases of read 1, respectively. The right mate of each read pair was trimmed to remove any portion of the SMART adaptor sequence or large stretches of poly(A) tails (6 consecutive bp or larger). The trimmed reads were then aligned to the human genome (GENCODE GRCH38) with STAR v2.5 with the default parameter settings. Reads that mapped to intronic or exonic regions of genes as per the GENCODE gene annotation were recorded. We applied one further correction step to fix barcode synthesis errors by inserting N at the last base of the cell barcode for reads in which the first 11 bases of the cell barcode were identical and the last T base of UMI was the same. Read-mapping statistics are listed in **Supplementary Table 1**. We calculated useful reads by adding together all uncollapsed mapped genic reads (generated by the Drop-seq pipeline) from each cell barcode that passed the filter (approximately equal to (Total raw reads) × (Proportion of reads containing poly(T) signal) × (Proportion of reads mapping to the genome as generated by STAR aligner) × (Proportion of reads mapping to genic (exon + intron) regions as generated from RSeQC) × (Proportion of reads associated with cell barcodes passing filter)). The digital expression matrix was then generated with genes as rows and cells as columns. We assigned UMI counts for each gene of each cell by collapsing UMI reads that had only 1 edit distance.

snDrop-seq data clustering and analyses. UMI matrix cell barcodes were tagged by their associated sequencing library batch ID (**Supplementary Table 1**) and combined across independent experiments. Mitochondrial genes not expressed in nuclei were excluded, and only UMI counts associated with protein-coding genes were used for clustering analyses. Nuclei with fewer than 300 molecules or more than 5,000 molecules (outliers) were omitted. We normalized molecular counts by using the total number of reads as the estimated library size for each cell. Variance normalization and clustering were done with the PAGODA2 package (https://github.com/hms-dbmi/pagoda2). Clustering and analysis were first performed separately for the visual cortex, frontal cortex, and cerebellum data sets. Briefly, the expression values were rescaled so that the mean expression of a gene in each measurement batch was equal to the data-set-wide average. We used a Winsorization procedure to cap the magnitude of the ten most extreme values for each gene. To estimate the residual variance for each gene, we modeled variance dependency on the expression magnitude (log scale) as a smoothed generalized additive model with smoothing term $k = 10$ (mgcv package in R). The observed-to-expected variance ratio for each gene was modeled by F distribution using the degrees of freedom corresponding to the number of successful gene observations. To normalize the contribution of each gene in the subsequent principal component analysis, we rescaled the variance of each gene to match the tail probability obtained from the F distribution under a standard normal sampling process. Cell clusters were determined from an approximate *k*-nearest-neighbors graph based on a cosine distance of the top 150 principal components derived from the top 2,000 variable genes from the variance-adjusted expression matrix, using the Infomap community detection algorithm (as implemented in the igraph R package). Cell clusters with fewer than 30 cells were omitted from further analysis. A preliminary round of clustering grouped low-depth cells that could not be confidently assigned to other clusters, and was omitted. Resulting cells were reclustered and visualized by *t*-distributed stochastic neighbor embedding (t-SNE) on the 150 principal components. Cell clusters were annotated manually on the basis of known markers for the frontal cortex, visual cortex, and cerebellum separately. For combined visualization, all data sets from the frontal cortex, visual cortex, and cerebellum were pooled and reclustered via the same general approach as described previously. The R script is provided at https://github.com/JEFworks/Supplementary-Code for additional information on parameters used for each individual and combined data set (Occ.R, Fcx.R, Cer.R, Combined.R).

We used Seurat software (V1.4.0.5) in R (https://github.com/satijalab/seurat) to construct violin plots and carry out differential gene expression analyses. For normalization, UMI counts for all annotated nuclei were scaled by the total UMI counts (excluding mitochondrial genes), multiplied by 10,000, and transformed to log space. Technical effects associated with UMI coverage and batch identity were regressed from scaled data with the RegressOut function in Seurat. Genes that were differentially expressed between cell types and subtypes were identified (Seurat software) by a likelihood-ratio test on all genes to identify 0.25-fold (log scale) enriched genes detected in at least 25% of cells in the cluster. Differential expression analyses were performed for all clusters, for excitatory or inhibitory neuron subtypes separately, for cerebellar data sets separately, or for all oligodendrocyte lineage cells separately (**Supplementary Table 3**).

Comparison of snDrop-seq data with published data. Control bulk RNA-seq data (FPKM values) from mouse cerebral cortex and human temporal lobe were obtained from http://web.stanford.edu/group/barres_lab/brain_rnaseq.html and http://web.stanford.edu/group/barres_lab/brainseqMariko/brainseq2.html, respectively. The top 50 cell-type-enriched genes were derived from comparison of averaged expression values of each cell type against an average of the remaining cell types (with the exception of oligodendrocyte subpopulations, which were compared only against non-oligodendrocyte lineages). Type-enriched genes from bulk data sets were used for correlation of log-averaged FPKM values of the associated bulk RNA-seq data with log-transformed average expression values from snDrop-seq data.

For comparison with single-cell RNA-seq data from human temporal lobe[28], we obtained gene count data from GEO (GSE67835) and normalized them with Seurat as described above, using a minimum cutoff of 1,000 genes detected. Highly variable genes were identified from a mean variability plot (average expression versus dispersion (variance/mean) assigned to 20 bins based on average expression) using a log(variance/mean) cutoff of 1 to identify 2,235 genes. Principal component analysis was performed on these highly variable genes, then projected to the entire data set. Statistically significant principal components ($P < 0.05$) were identified by a jackstraw approach. Cell identities from the original publication were maintained, and the top 50 genes from the statistically significant principal components differentiating these cell types, as well as the top 10 differentially expressed genes associated with each cell type, were identified by Seurat and used for correlation of log-transformed averaged expression values from scRNA-seq and snDrop-seq data. For comparison with scRNA-seq data from mouse visual cortex[27], gene RPKM data was obtained (GSE71585), log-transformed, and loaded into Seurat with published cluster annotations. Neuronal subtypes were combined into a single group and average cluster expression values were obtained across cell types, using previously described marker genes present in each cluster[27], and a correlation heat map of log-transformed averaged expression values was generated. SNS data generated on the Fluidigm C1 platform[4] (dbGaP accession phs000833.v3.p1) were used for correlation of log-transformed subtype-averaged expression values for differentially expressed genes (greater than twofold) underlying previous subtype clustering and classifications[4]. For pairwise sample correlations, all differentially expressed genes (greater than twofold) identified during clustering of all data sets were used.

For comparison of UMI counts and genes detected with scDrop-seq data from mouse retina[17], the full UMI count table for 44,808 annotated samples was obtained from GEO (GSE63472). For comparison with 9k brain cell data sets from an E18 mouse generated on the 10X platform (Cell Ranger 1.3., v2 Chemistry), filtered gene matrices were downloaded from the company website (https://www.10xgenomics.com). For comparison with human embryonic midbrain single-cell data sets generated with the Fluidigm C1 platform[60], annotated UMI count matrices were obtained from GEO (GSE76381). Each data set was analyzed with Seurat for t-SNE visualization of clusters.

RNA *in situ* hybridization and protein expression data. Combinatorial RNA *in situ* hybridization experiments (**Fig. 2e**,**h**, **Supplementary Figs. 8d** and **9d**) were carried out with the RNAscope multiplex fluorescence kit (*SLC17A7*, *EYA4*, *GAD1*, *SORCS3*) or the RNAscope brown chromogenic kit (*CBLN*, *PCP4*) according to the manufacturer's instructions (Advanced Cell Diagnostics) and as previously described[4] and outlined in **Supplementary Table 11**. RNAscope counts were obtained for four separate layer cross-sections (replicate regions) (**Supplementary Table 11**), and averaged values and s.d. (error bars) were plotted. For improved visualization of the *GAD1*/*SORCS3* stains shown in **Figure 2h**, images were further adjusted for contrast in ImageJ; however, counts were performed on representative images shown in **Supplementary Figure 9d**. For single-gene RNA *in situ* hybridization from visual or frontal cortex (**Fig. 2c**, **Supplementary Fig. 8e**), representative images were obtained from the Allen Human Brain Atlas (http://human.brain-map.org); corresponding links are provided in **Supplementary Table 9**. For individual protein stains (**Fig. 2g**, **Supplementary Fig. 9c**), representative images were obtained from The Human Protein Atlas (http://www.proteinatlas.org) and are referenced in **Supplementary Table 10**. All image panels were assembled in Adobe Illustrator and/or Adobe Photoshop.

scTHS-seq sample origin and nuclei preparation. The human tissue samples used for each scTHS-seq experiment are listed in **Supplementary Table 1**. After flow cytometry, nuclei were kept on ice and spun down at 500*g* for 5 min at 4 °C, after which supernatant was removed and the pellet was resuspended in 1× lysis buffer (1× concentration: 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 2% BSA, one Roche protease inhibitor tablet per 10 mL, in PBS) and chilled at 4 °C for 5 min without shaking. Then nuclei were spun down at 500*g* for 5 min at 4 °C, supernatant was removed, and the pellet was resuspended in 1.5× tagmentation buffer (1.0× concentration: 33 mM Tris-OAc, pH 7.8, 66 mM K-OAc, 10 mM Mg-OAc, 16% dimethylformamide). At that point the nuclei sample was ready for nuclei counting and species–species sample mixing. For scTHS-seq, a mouse nuclei sample and a human nuclei sample were always mixed so we could perform assay quality control and calculate collision rates, ensuring low collision rates were achieved. This is discussed further in the subsection "scTHS-seq collision rate determination." The mouse nuclei sample was prepared the same way after flow cytometry. For species–species mixing, both human nuclei and mouse nuclei samples were counted on a Bio-Rad TC20 cell counter and diluted with 1.5× tagmentation buffer, or were further concentrated by being spun down at 500*g* for 5 min at 4 °C and resuspended in a smaller volume with 1.5× tagmentation buffer, so the samples' cell counts were within 10% of each other. A cell concentration of ~2.4×10^5 nuclei/mL was obtained for each sample, with the optimal range being 2.0×10^5 to 5.0×10^5 nuclei/mL with ~1 million total nuclei for each sample. Next, equal volumes of mouse and human samples were combined and mixed gently. The sample was then ready for transposition and combinatorial indexing.

scTHS-seq transposon generation. Each transposon consisted of two oligos synthesized by IDT and kept as 100 µM stock solutions in TE buffer: the 74-bp barcoded transposon and the 19-bp universal 5′ phosphorylated mosaic end. In total, there were 384 barcoded r5 transposons, each with a unique 6-bp barcode, and all barcodes had a minimum edit distance of 2 (**Supplementary Table 12**). For the generation of annealed transposons, 10 µL of each 100 µM oligo was added to each well of a 384-well plate (final concentration: 50 µM), incubated at 95 °C for 2 min, cooled to 14 °C at 0.1 °C/s, diluted to 8.4 µM in TE buffer with a final concentration of 50% glycerol, and then stored at −20 °C.

Generation of scTHS-seq barcoded transposome complex. Tn5059 was generated and normalized for activity at Illumina. The mutations and methods of protein expression and purification used to independently generate Tn5059 have been published[19]. Because complexed Tn5059 and transposons slowly lose activity over time, with a noticeable loss in data quality after a few weeks, r5 transposome complexes were generated freshly for each scTHS-seq run and used within a few days. First, Tn5059 was diluted to 4.2 µM in standard storage buffer (Illumina), and 1 µL was added to each well of 384-well plate. Next, 1 µL of 8.4 µM annealed barcoded r5 transposon was added to each well, and the 384-well plate was incubated at room temperature for 30 min. For custom nXTv2_i7 Tn5059 transposome generation, the annealed nXTv2_i7 transposon (50 µM) was generated as described in the subsection "scTHS-seq transposon generation" (**Supplementary Table 12**). To generate a complexed transposome solution, we incubated 7 µM Tn5059 with 10 µM annealed transposon for 30 min at room temperature and then diluted to 0.7 µM Tn5059 transposome complex with standard storage buffer (Illumina). These custom i7 transposome complexes were stored at −20 °C and used within a few days.

scTHS-seq nuclei tagmentation and barcoding. To the 384-well plate of freshly generated uniquely barcoded Tn5059 r5 transposome complexes, we added 4 µL of human–mouse mixed cell sample for a total of ~960 nuclei per well (optimally ~2,000 nuclei per well) and a final concentration of 0.7 µM Tn5059 r5 transposome complex. Each sample was mixed gently five times with an electronic pipettor and incubated at 37 °C for 30 min. To stop the reaction, we added 4.0 µL of 50 mM EDTA to each well and mixed gently five times with the electronic pipettor, and then incubated the mixture at 37 °C for 15 min before storing it at −20 °C overnight. The next day, samples were thawed, one volume of cold 2× FACS buffer (1× FACS buffer: 2 mM EDTA, 1% BSA in PBS) was added to each well, and samples were mixed gently three times with the electronic pipettor and pooled into one tube on ice. That tube was then spun down at 500*g* for 5 min at 4 °C. Supernatant was removed, and tagmented nuclei were resuspended in 1.5 mL of cold 1× FACS buffer. Next, 75 µL of propidium iodide (PI; eBioscience) was added, and nuclei were sorted by flow cytometry into 96-well plates containing 10 µL of PBS per well at 100 nuclei per well and kept on ice. Doublets were removed on the basis of forward and side scatter plots, and PI-staining events were selected.

scTHS-seq library preparation. Each 96-well plate of nuclei was processed individually. First, 11 µL of guanidine hydrochloride was added to each well and mixed by light vortexing. Reactions were purified by the addition of 40 µL of (1.8×) AMPure SPRI beads followed by light vortexing, and then bead-pelleting and 80% ethanol washes were performed with the 'flick and blot' method and a magnetic plate from V&P Scientific. After 80% ethanol washes were complete, the plate was quickly spun down at 500*g*, and leftover

doi:10.1038/nbt.4038

80% ethanol was removed by pipetting. 10 µL of 1× NEB Taq polymerase was added to each reaction, and the plate was lightly vortexed to resuspend the beads (SPRI beads left in the reaction), after which the reactions were run at 72 °C for 3 min for end fill-in and then placed on ice. For *in vitro* transcription amplification, we used the NEB HiScribe T7 high-yield synthesis kit. We added a master mix of 2 µL of 10× transcription buffer, 2 µL of ATP, 2 µL of CTP, 2 µL of GTP, and 2 µL of UTP directly to the end fill-in reactions, and then each reaction was lightly vortexed and subsequently incubated at 37 °C for 19 h. After incubation, a couple of samples were run on a TBU gel to verify that *in vitro* transcription amplification had occurred. Reactions were purified by the addition of 44 µL of (2.0×) SPRI binding buffer (20% PEG 8000, 2.5 M NaCl, 10 mM Tris-HCl, 1 mM EDTA) to each reaction, and the plate was vortexed thoroughly. 80% ethanol washes and removal of leftover 80% ethanol were performed as described above, and SPRI beads were resuspended in 9 µL of nuclease-free water. For reverse transcription, we added 2.5 µL of 20 µM random hexamers to each reaction, then vortexed the plate lightly and heated it to 70 °C for 3 min, after which we immediately cooled it on ice. We used the Clontech SMART MMLV reverse transcriptase kit with the addition of 4 µL of 5× first-strand synthesis, 2 µL of dNTP mix, 2 µL of 100 mM DTT, and 0.5 µL of SMART MMLV RT in a master mix to each reaction and vortexed the plate lightly. Reactions were incubated at 22 °C for 10 min, then 42 °C for 60 min, and terminated at 70 °C for 10 min. To degrade RNA in cDNA–RNA hybrids, we added 1 µL of 0.5 units Enzymatics RNase H to each reaction, vortexed the plate lightly, and incubated the plate at 37 °C for 20 min. For second-strand synthesis, we added the first 2.5 µL of 20 µM sss_scnXTv2 (**Supplementary Table 12**) to each reaction and lightly vortexed it, then incubated it for 2 min at 65 °C and immediately cooled it on ice. Then we added 5.9 µL of NEB taq5X to each reaction and incubated it at 72 °C for 8 min. After the samples had cooled on ice, we added 60 µL of (2.0×) SPRI binding buffer to each sample. The plate was vortexed thoroughly and subjected to 80% ethanol washes and removal of leftover 80% ethanol as described previously, and SPRI beads were resuspended by light vortexing in 7 µL of nuclease-free water. Double-stranded cDNA fragments then underwent simultaneous fragmentation and 3′ adaptor addition with a custom nXTv2_i7 Tn5059 transposome (**Supplementary Table 12**). To 7-µL volumes of each sample, we added 2 µL of 5× tagmentation buffer, followed by 2 µL of prepared 0.7 µM custom nXTv2_i7 Tn5059 transposomes (final transposome concentration of 0.14 µM). The sample was then vortexed lightly, incubated at 55 °C for 6 min, and cooled briefly on ice. Immediately after the cooling period, we added 19 µL of 6.32 M guanidine hydrochloride, for a final guanidine hydrochloride concentration of 4 M, to each reaction and briefly vortexed the sample. Then 60 µL of (2.0×) SPRI binding buffer was added to each sample. The plate was vortexed thoroughly, and 80% ethanol washes and removal of leftover 80% ethanol were performed as described previously. However, for this purification SPRI beads were resuspended in 16 µL of nuclease-free water and the plate was placed back on the magnetic plate. We eluted sample off SPRI beads held by the magnetic plate and transferred it to a qPCR plate. Standard Illumina Nextera XT v2 barcoding in an 8 × 12 (i5 × i7) format was performed with qPCR, using custom scTHS-seq i5 indexes and standard Illumina i7 indexes (**Supplementary Table 12**). In total, 20 µL of KAPA SYBR Fast, 2 µL of 10 µM scT7_S5XX index primer, and 2 µL of 10 µM nXTv2_i7XX index primer were added to each reaction, for a total volume of 40 µL, and mixed well. qPCR was run at 72 °C for 3 min and then 95 °C for 30 s, and cycled at (95 °C for 10 s, 63 °C for 30 s, 72 °C for 1 min) until curves reach saturation—typically 9–12 cycles. Plates were stored at −20 °C.

scTHS-seq library validation, pooling, and sequencing. To validate libraries, we ran 1 µL of each qPCR reaction on 6% Tris-borate-EDTA (TBE) gels stained with SYBR Gold. For pooling, 2 µL (4 µL or 6 µL if yields were low) of each uniquely barcoded qPCR reaction was combined and size-selection was performed as described[9]. Resultant size-selected libraries were quantified with Qubit and sequenced on an Illumina MiSeq system (50 + 32 + 32 single-end reads) for validation, then on the high-throughput Illumina HiSeq 2500 (50 + 8 + 32 single-end reads) for data generation.

scTHS-seq data processing. Raw BCL files were demultiplexed to FASTQ Read1, Index1, and Index2 files with bcl2fastq v2.17.1.14, then used as input to deindexer (https://github.com/ws6/deindexer) with 0 mismatch barcode demultiplexing. Barcode combinations associated with each read were appended to each read header with in-house Perl scripts, and all FASTQ files were combined and mapped to an hg38 no-alternative loci plus decoy reference genome (GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set) and mm10 no-alternative loci reference genome (GCA_000001635.5_GRCm38.p3_no_alt_analysis_set) using BWA 0.7.12-r1039. Mapped SAM outputs were re-demultiplexed by barcode and converted to BAM files, and clonal reads were removed with SAMtools 1.3.1, while read statistics were gathered for each barcode combination. To determine which uniquely barcoded nuclei were suitable for downstream analyses, we filtered nuclei requiring $\log_{10}$(Total reads + 1) > 3.

Joint peak calling was performed on pooled BAM files with SPP (v1.13; https://github.com/hms-dbmi/spp). In total, 32,869 cells (**Supplementary Table 2**) were pooled. Reads that mapped within 100 bp of known repeat regions according to annotations from Repeat Masker (http://www.repeatmasker.org) were removed. A smoothed density of pooled reads was generated from window tag counts with a window size of 500 bp and a window step of 100 bp. DNA-accessible regions (peaks) were called on the basis of the smoothed density with a minimum threshold of five reads and minimum span of five steps between each peak. Peaks were filtered using a permutation-based false discovery rate of 10×10^{-8} and filtered for presence in at least 30 cells from the visual cortex, which resulted in 52,694 final peaks called. Called peaks were then assessed for reads in each individual cell to generate a matrix of peaks versus cells for downstream clustering and analysis. The R script used (spp_comb.R) is provided at https://github.com/JEFworks/Supplementary-Code.

scTHS-seq data clustering and analysis. The molecular count matrix was binarized for further analysis. We selected 52,694 sites that were observed in 30 or more cells of the visual cortex. Variance normalization and clustering were done with a modified model in the PAGODA2 package to better represent the limited dynamic range of scTHS-seq data. Clustering and analysis were carried out separately for the visual cortex, frontal cortex, and cerebellum data sets. Briefly, data were modeled as a right-censored Poisson process (observing at most one molecule per site). To determine cell depth and batch-specific site observation probabilities, we used an expectation-maximization algorithm, with each iteration fitting MLE values for library size and batch-specific site probabilities sequentially. To evaluate overdispersion of each site, we calculated the total deviance across all observations for a given site under the censored Poisson process. The relationship between the total deviance and mean site occurrence frequencies was modeled with a generalized additive model (mgcv R package; smoothing term $k = 10$). The observed/expected deviance difference was scored with a variance gamma distribution. To cluster and visualize the cells in the visual cortex, we determined the top 30 principal components on the censored Poisson deviance residual matrix. The negative deviance residuals associated with non-observed sites were ignored. Cell clusters were determined on a *k*-nearest-neighbors graph ($k = 50$) using a multilevel community detection method (igraph R package). Cell clusters with fewer than 30 cells were omitted from further analysis. We achieved two-dimensional visualization by applying t-SNE to the 30 principal components, using a perplexity of 50. The 30 principal components derived from the visual cortex were also used to project cells from the frontal cortex and cerebellum. Because there were fewer cells in the frontal cortex and cerebellum data sets than in the visual cortex data set, a smaller *k* value of 30 and perplexity of 30 were used for the *k*-nearest-neighbors graph and t-SNE embedding, respectively. Cell clusters were annotated on the basis of the accessibility of marker genes, as well as the scTHS-seq and snDrop-Seq joint analysis described below. Upon inspection, the three smallest clusters seemed to represent poorly resolved cells or doublets mixing signals from two or more subpopulations (based on artificial mixing of cells from other populations) and were annotated or examined in further analysis. The R scripts used for analysis of the visual cortex (scTHSSeq_Occ.R) and projection of other data sets (scTHSSeq_other.R) are provided at https://github.com/JEFworks/Supplementary-Code.

scTHS-seq collision-rate determination. For each unique barcode combination, the proportion of unique reads that mapped to either mouse or human genome was calculated (**Supplementary Table 1**). A unique barcode combination was determined to belong to one species if 89% of the reads mapped to one genome; otherwise the barcode combination was determined to be a detectable collision. This calculation excludes the possibility of three nuclei collisions, which would represent extremely rare events. For visualization, results were then graphed in R in an *x*,*y* scatter plot and density plot.

scTHS-seq and snDrop-seq joint analysis. To map between transcriptional and epigenetic space, we trained a gradient-boosted regression model (GBM) to predict the probability of differential expression from patterns of nearby accessibility differences, and a separate GBM in reverse to predict the probability of differential accessibility given the differential expression observations. The GBM was implemented with the caret (V6.0-72) package in R. The prediction GBM used the following features: mean expression of the associated gene; distance of the site to the gene's transcription start site; differential expression *Z*-score of the gene; fold enrichment of the gene; Boolean representations of whether the site is in a promoter, exon, distal intergenic region, 5′ UTR, genic region, intergenic region, region immediately downstream of the gene end, intron, or 3′ UTR; and whether the gene showed the highest expression in one cluster compared with all the others. Models were trained on astrocyte and oligodendrocyte data from the visual cortex only to learn the relevance of features as weights (**Supplementary Table 13**). Models were fit by 10× cross-validation. Joint scores (across multiple genes or sites) were calculated as probability means of individual elements (sites or genes).

We applied our classifier to identify epigenetic subpopulations from our scTHS-seq data, integrating information from the finer-resolution snDrop-seq data. To do this, we first carried out hierarchical clustering on cell-type similarities based on the expression of all genes to establish a cell-type relationship dendrogram. We then iteratively performed binary splits on this dendrogram and identified significantly differentially upregulated genes (Z-score > 1.28) in each branch by Fisher's exact test. We applied our GBM to predict differentially accessible genomic sites.

To classify scTHS-seq cells as corresponding to either branch, we assessed accessibility in the predicted accessible sites for each branch, normalized by the number of accessible sites observed in total for each cell. Thus, cells with high accessibility of sites predicted to be accessible in branch A were assigned as such. Ties were randomly broken. Having identified putative corresponding subpopulations in scTHS-seq data, we refined the predicted branch annotations by identifying significantly differentially accessible sites (Z-score > 1.28) by Fisher's exact test, and reassessed each cell's joint accessibility. Refinement was repeated until convergence, that is, until cell branch annotations no longer changed by more than 10%. This typically required 2–5 repeats. Finally, we assessed the stability of the branch annotations by using, randomly, 90% of cells from each group to identify differentially accessible sites that were used to derive joint accessibility scores for the remaining 10% of cells. Stability was quantified as the area under the ROC curve from joint accessibility scores with the original annotations.

To enhance the separation of refined subpopulations in our data visualization, we identified differentially accessible sites for each refined subpopulation and computed the joint accessibility scores for each cell and each refined subpopulation. We applied t-SNE to the joint accessibility scores in addition to the original 30 principal components to achieve a refined 2D embedding that better segregated our refined subpopulations (**Fig. 3e**,**f**).

scTHS-seq transcription factor analysis. To infer relevant TFs and transcription factor binding sites (TFBSs), we obtained DNA sequences corresponding to scTHS-seq peaks and position weight matrices (PWMs) for 379 TFs from the JASPAR database. A sliding window was used to identify the maximum PWM score for each peak, taking into consideration both plus and minus strands. We normalized PWM scores within each peak by subtracting the theoretical minimum and dividing by the maximum score possible for each PWM using the PWMscoreStartingAt() function from the matchPWM package in R, assuming a uniform prior distribution of all nucleotides. We then standardized scores for each peak and TF to *Z*-scores by subtracting the mean and dividing by the s.d. of scores for each TF to control for background rates of binding and nonspecificity. TFBSs were inferred as corresponding to peaks with Z-score > 1.96 for each TF. We assessed the overlap of inferred TFBSs with previously identified cell-type-specific peaks by Fisher's exact test. TFs with TFBSs that significantly overlapped cell-type-specific peaks (Bonferroni-corrected $P < 0.2$) were inferred to be relevant to the cell type. We integrated snDrop-seq data to assess the expression fold change of these TFs in each cell type, assessing significance by using rank-based gene set enrichment analysis. Specifically, TF expression was averaged across cells for each cell type. A $\log_2$ fold change comparing the average expression in oligodendrocytes versus neuronal cell types was used to assess the enrichment of expression for predicted oligodendrocyte-related TFs. Gene set enrichment analysis was carried out with the LIGER package in R (https://github.com/JEFworks/liger).

scTHS-seq GWAS data analysis. GWAS SNPs were downloaded from the GRASP database, using categories with any trait for selection with $P < 1 \times 10^{-6}$. The selected categories were Alzheimer's disease, schizophrenia, Parkinson's disease, bipolar disorder, autism, multiple sclerosis, attention deficit hyperactivity disorder, amyotrophic lateral sclerosis, epilepsy, depression, glaucoma, Crohn's disease, celiac disease, type 1 diabetes, lung disease, chronic kidney disease, and prostate cancer. For the rest of the analysis, in-house Python and shell scripts were used. For each category, all SNPs were extended at each end to encompass a 100-kb region. Any SNP regions that overlapped were merged in bedtools to generate a larger SNP region containing both SNPs. Next, for each SNP region, the SNP with the most significant *P* value was selected. This removed any instances of multiple linked variants for the same trait, and ensured that there were no variants in linkage disequilibrium. Next, the SNPs with the top 50 most significant *P* values and their gene regions were selected for further analysis. To determine the overlap of accessible regions in each cluster defined during cell clustering and identification, we carried out peak calling with SPP v1.2 using the merged data of each cluster to generate a list of peak regions, and then we generated lists of differential peaks (peaks present in one or more cell types but not others) for all the cell types. Peaks with Z-scores < 400 were removed to generate a final peak list for each cluster. Next, those peaks were overlapped with the top 50 SNP regions for each disease category, and the number of overlaps was counted. To determine whether enrichment was significant, we calculated *Z*-scores. First, we carried out 20,000 permutations of the peak regions over the hg38 reference genome using only autosomes, and for each permutation we counted overlaps of peaks for each cluster with SNP regions. From the permutations, averages and s.d. were calculated, and in conjunction with previously calculated total overlaps, the *Z*-score for each cluster was calculated. For visualization, R was used to overlap *Z*-scores onto the clusters and generate a heat map of the similarity between cell types and diseases. For the excitatory and inhibitory subclusters, we carried out the same analysis with slight modifications: all peaks were kept for analysis (instead of removing peaks with Z-scores < 400), because there were fewer differential peaks overall between the subclusters. This was due to the exclusion of differential peaks that would define the main excitatory and inhibitory clusters and that were not differential between the subclusters.

Bulk ATAC-seq microglia data set comparison. Raw bulk microglia ATAC-seq FASTQ files were obtained from Gosselin *et al.*[53] and mapped with BWA 0.7.12-r1039 to Hg38. Peak-calling was performed with Dfilter 1.0, and peaks were overlapped with the differential peak files for each cluster from visual cortex data. For GWAS risk variant enrichment analysis, the peak file was run though the same pipeline with the same parameters as the visual cortex differential peak cluster files.

Developmental ordering of oligodendrocyte lineage data sets. To order cells according to their developmental trajectory along the oligodendrocyte lineage, we selected 3,064 snDrop-seq data sets for cells from the visual cortex identified as OPCs (644 data sets) or Oli (2,420 data sets) by the previous PAGODA2 clustering-based approaches. A diffusion map approach applied with the Destiny package[38] in R was applied to normalized counts with parameter $k = 100$ and otherwise default parameters. Cells were ordered according to their value along the first eigenvector. To identify OPC, immature Oli, and mature Oli genes along the developmental trajectory, we selected the first 400 cells as representative OPCs, the 700th to 1,100th cells as representative immature Oli, and the 2,664th to 3,064th cells as representative mature Oli. Differentially upregulated genes from each group were identified with PAGODA2. Gene Ontology annotations for each gene set (**Supplementary**

doi:10.1038/nbt.4038

Table 6) were obtained from ToppGene (https://toppgene.cchmc.org). To establish a corresponding trajectory according to accessibility, we selected 5,077 scTHS-seq data sets for cells from the visual cortex as identified as OPC (505 data sets) or Oli (4,572 data sets) by the previous PAGODA2 clustering-based approaches. All peaks were annotated using the ChIPseeker package[61] with annotations from the TxDb.Hsapiens.UCSC.hg38.knownGene package. For differentially upregulated genes from each group, joint accessibility was quantified as the average accessibility of all sites corresponding to said genes multiplied by 1×10^6. In this manner, a joint accessibility score was derived from each cell for OPC, immature Oli, and mature Oli genes. Joint accessibility scores were scaled and clustered with hierarchical clustering and a ward.D2 linkage for visualization. The R script used (Destiny.R) is provided at https://github.com/JEFworks/Supplementary-Code.

As a separate supportive trajectory analysis of expression data, we analyzed all OPC and Oli snDrop-seq data sets generated across all regions (**Supplementary Table 2**) with Monocle (v2.4.0)[62] according to the provided documentation (http://cole-trapnell-lab.github.io/monocle-release/) and with the following parameters: UMI counts were modeled as a negative binomial distribution; ordering genes were identified as having high dispersion across cells (mean_expression >= 0.005; dispersion_empirical >= 1); and the discriminative dimensionality reduction with trees (DDRTree) method was used to reduce data to two dimensions. Gene sets identified from the Destiny analysis were clustered and visualized using the plot_pseudotime_heatmap function and gene ontologies identified from https://toppgene.cchmc.org/.

Statistics. Combined snDrop-seq analyses were performed on 35,442 single-nucleus data sets generated over 20 experiments, each split into 1–6 libraries, for 46 libraries in total (**Supplementary Table 1**). For brain regions analyzed, biological replicates included visual cortex (five individuals), frontal cortex (four individuals), and cerebellar hemisphere (four individuals). For analyses on individual regions, 19,368 (visual cortex), 10,319 (frontal cortex), and 5,602 (cerebellar hemisphere) single-nucleus data sets were used (**Supplementary Table 2**). Differential expression of genes between cell-type clusters (the number of data sets per cluster is listed in **Supplementary Table 2**) were determined with the "bimod" likelihood-ratio test in Seurat; *P* values and false discovery rates (FDR < 0.05) are listed in **Supplementary Table 3**.

To identify gene expression signatures associated with remyelination in the visual cortex, we carried out differential expression analysis of a limited set of 400 OPCs, 400 immature Oli, and 400 mature Oli identified on the basis of pseudotime ordering from the Destiny analysis (**Supplementary Table 6**).

For scTHS-seq analyses, 32,869 single-cell data sets were generated from three experiments, each split into two (visual and frontal cortex) or three (cerebellum) libraries for sequencing. For each region, data sets (13,232 visual cortex; 4,753 frontal cortex; 9,921 cerebellar hemisphere) were generated from a single individual, with a different individual used for each region, for a total of three individuals (**Supplementary Table 2**). For analyses across regions, 15,786 combined data sets were used (**Supplementary Table 2**).

To identify potentially important cell-type-specific TFs from scTHS-seq data, we screened a set of 379 TFs with known PWMs from the JASPAR database for significant over-representation within differentially accessible peaks associated with each cell type (Ex versus In versus End versus Ast versus Oli versus OPC versus Mic) or cell-type subpopulation (ExL23 versus ExL4 versus ExL56, InA versus InB, OPC versus immature Oli versus mature Oli) in the visual cortex (the number of data sets per group is listed in **Supplementary Table 2**). The significance of over-representation was assessed by Fisher's exact test (n = 13,232 data sets total) with Bonferonni multiple-testing correction (**Supplementary Tables 5** and **7**).

To identify cell-type-specific risk variant enrichments for common genetic diseases, we defined the top 50 most significant SNP regions for each disease using SNPs from the GRASP database. For each cell type within the visual cortex (the number of data sets per group is listed in **Supplementary Table 2**), a list of differential peaks was defined (peaks present in one or more cell types and not others, only peaks with Z-score > 400). For Z-score determination, differential peaks from each cell type were overlapped with the top 50 SNP regions for a disease, and the number of overlaps was counted. Next, we carried out 20,000 permutations of the peak regions on the hg38 reference genome using only autosomes, and counted overlaps in the top 50 SNP regions for each permutation. From this, averages and s.d. were calculated, and in conjunction with previously calculated total overlaps, the Z-score for each cell type was calculated. For the excitatory and inhibitory subclusters, the same analysis was performed, with the exception that all peaks were kept for analysis (instead of removing peaks with Z-scores < 400). For the published bulk microglia ATAC-seq data, the same analysis was performed, with the exception that all peaks were kept for analysis. Z-scores are listed in **Supplementary Table 8**.

Life Sciences Reporting Summary. Further information about experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. Raw sequencing data, annotated snDrop-seq and scTHS-seq count matrices, and DNA accessibility peak files are all available from the Gene Expression Omnibus under SuperSeries accession code GSE97942.

60. La Manno, G. *et al.* Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167**, 566–580 (2016).
61. Yu, G., Wang, L.G. & He, Q.Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
62. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

nature research

Corresponding author(s): J. Chun, P. Kharchenko, K. Zhang

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.	snDrop-seq was performed at least 2-times per individual and per region (6 individuals total). scTHSseq was performed for a single individual for each region, with each run capable of generating sufficient sampling. Final dataset number was determined by quality filtering metrics described in the methods

2. Data exclusions

Describe any data exclusions.	Low quality data sets were excluded based on parameters outlined in the methods section (e.g. low read mapping).

3. Replication

Describe whether the experimental findings were reliably reproduced.	All attempts at replication were successful

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	All data sets for each brain region (snDrop-seq or scTHS-seq) were combined and clustered using an unbiased approach

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	Individual data set clustering was performed using an unbiased and unsupervised approach prior to cell type assignments

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Biotechnology: doi:10.1038/nbt.4038

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	All software used in this study has been described in published literature (e.g. Seurat, Destiny, Monocle, etc...) or are custom and available on Github (e.g. Pagoda2). Custom code needed to repeat analyses has also been made available on Github, as documented in the Methods section.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	The Tn5059 hyperactive transposase was acquired through a collaboration with Illumina, however, the mutations and enzyme production methods needed to generate this enzyme have been published and are referenced in the Methods. All other reagents are commercially available

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	All antibody stains were obtained from the human protein atlas (http://www.proteinatlas.org) which provides antibody validation and availability information

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	293T and NIH/3T3 cell lines (species mixing snDrop-seq experiment) were from ATCC
b. Describe the method of cell line authentication used.	Genomic mapping of RNA-seq reads
c. Report whether the cell lines were tested for mycoplasma contamination.	Cell lines were not tested for mycoplasma contamination
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	No commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used in this study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No human research participants were used in this study